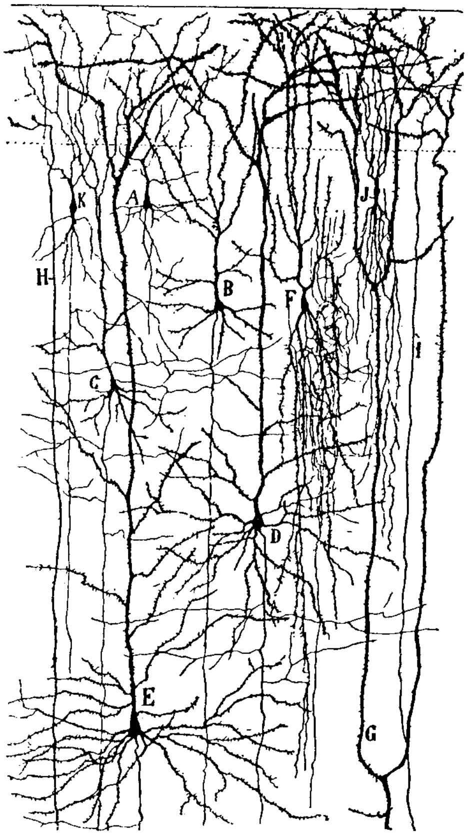
Statistical Mechanics of Neural Networks

Lecture Notes of Course G32/NN13

August 1997



A.C.C. Coolen
Department of Mathematics
King's College London

Contents

Preface

Statistical mechanics deals with large systems of stochastically interacting microscopic elements (particles, magnets, polymers, etc.). The general strategy of statistical mechanics is to abandon any ambition to solve models of such systems at the microscopic level of individual elements, but to use the microscopic laws to calculate laws describing the behaviour of a suitably choosen set of *macroscopic* observables. The toolbox of statistical mechanics consists of various methods and tricks to perform this reduction from the microscopic to a macroscopic level, which are based on clever ways to do the bookkeeping of probabilities. The experience and intuition that has been built up over the last century tells us what to expect (e.g. phase transitions), and serves as a guide in choosing the macroscopic observables and in seeing the difference between relevant mathematical subtleties and irrelevant ones. As in any statistical theory, clean and transparent mathematical laws can be expected to emerge only for large (preferably infinitely large) systems.

Neuronal (operation) processes as well as synaptic (learning) processes in neural networks appear to meet the criteria for statistical mechanics to apply, provided we are happy to restrict ourselves to (preferably infinitely) large systems. Here the microscopic stochastic dynamical variables are the states of the neurons (in the case of operation) or the values of the synapses (in the case of learning processes), and one is usually as little interested in knowing all individual neuron states or synapses as one would be in knowing all position coordinates of the molecules in a bucket of water. We are rather after overall performance measures such as pattern retrieval quality (in the case of associative memories) or generalization and training errors (in the case of networks learning to perform a given rule), which are indeed *macroscopic* observables.

Chapter 1

Microscopic Dynamics

1.1 Stochastic Local Field Alignment

Parallel Dynamics. The microscopic laws governing the parallel evolution of a system of N Ising spin neurons $\sigma_i \in \{-1,1\}$ are defined as a stochastic alignment to local fields $h_i(\boldsymbol{\sigma})$. These fields represent the post-synaptic potentials of the neurons and are assumed to depend linearly on the instantaneous neuron states:

$$\sigma_i(t+1) = \ \mathrm{sgn}\left[\tanh\left[\beta h_i(\boldsymbol{\sigma}(t))\right] + \eta_i(t)\right] \tag{1.1}$$

$$h_i(\boldsymbol{\sigma}(t)) \equiv \sum_{j=1}^{N} J_{ij}\sigma_j(t) + \theta_i(t) \tag{1.2}$$

The stochasticity is in the independent random numbers $\eta_i(t)$ (representing threshold noise), which are distributed uniformly over the interval $[-1,1]$. The parameter β controls the impact of this noise on the states $\sigma_i(t+1)$. For $\beta = \infty$ the random numbers cannot influence the system state and the process becomes deterministic: $\sigma_i(t+1) = \ \mathrm{sgn}[h_i(\boldsymbol{\sigma}(t))]$. The opposite extreme is choosing $\beta = 0$, in which case the system evolution becomes fully random. The external fields $\theta_i(t)$ represent neural thresholds and/or external stimuli. The specific choice *tanh* for the non-linearity in definition (1.1) is only relevant for the special case of symmetric interactions. There it allows us to identify the sequential version of this stochastic dynamics as a Glauber (1963) dynamics with respect to the standard Ising spin Hamiltonian, and as a consequence apply standard equilibrium statistical mechanics.

The microscopic equations (1.1) can be transformed directly into equations for the evolution of the microscopic state probability $p_t(\boldsymbol{\sigma})$, with $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$. If $\boldsymbol{\sigma}(t)$ is given we find

$$p_{t+1}(\boldsymbol{\sigma}) = \prod_{i=1}^{N} \frac{1}{2}\left[1 + \sigma_i \tanh[\beta h_i(\boldsymbol{\sigma}(t))]\right] = \prod_{i=1}^{N} \frac{e^{\beta\sigma_i h_i(\boldsymbol{\sigma}(t))}}{2\cosh[\beta h_i(\boldsymbol{\sigma}(t))]}$$

If, instead of $\boldsymbol{\sigma}(t)$, the probability distribution $p_t(\boldsymbol{\sigma})$ is given, the above expression generalises to the corresponding average over the states at time t:

$$p_{t+1}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_t(\boldsymbol{\sigma}') \tag{1.3}$$

$$W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\equiv\prod_{i=1}^{N}\frac{e^{\beta\sigma_i h_i(\boldsymbol{\sigma}')}}{2\cosh[\beta h_i(\boldsymbol{\sigma}')]} \tag{1.4}$$

which is the Markov equation corresponding to the parallel process (1.1).

Sequential Dynamics. In the case of sequential dynamics we assume that at each iteration step only one single spin σ_{ℓ_t} (to be drawn at random) will be updated:

$$\begin{aligned} i\neq\ell_t: &\quad \sigma_i(t+1)\equiv\sigma_i(t) \\ i=\ell_t: &\quad \sigma_i(t+1)\equiv\ \mathrm{sgn}\left[\tanh\left[\beta h_i(\boldsymbol{\sigma}(t))\right]+\eta_i(t)\right] \end{aligned} \tag{1.5}$$

with the local fields defined by (1.2). The stochasticity is now both in the independent random numbers $\eta_i(t)$ (the noise), which are distributed uniformly over the interval $[-1,1]$, and in the site ℓ_t to be updated, which is drawn randomly from the set $\{1,\ldots,N\}$ of all sites.

The microscopic equations (1.5) can again be transformed into equations describing the evolution of the microscopic state probability $p_t(\boldsymbol{\sigma})$. If $\boldsymbol{\sigma}(t)$ and the site ℓ_t to be updated are given we find

$$\mathrm{Prob}\left[\sigma_\ell(t+1)\right]=\frac{1}{2}\left[1+\sigma_\ell(t+1)\tanh[\beta h_\ell(\boldsymbol{\sigma}(t))]\right]$$

After averaging over the random site ℓ_t we obtain:

$$p_{t+1}(\boldsymbol{\sigma})=\frac{1}{N}\sum_{i=1}^{N}\left\{\left[\prod_{j\neq i}\delta_{\sigma_j,\sigma_j(t)}\right]\frac{1}{2}\left[1+\sigma_i\tanh[\beta h_i(\boldsymbol{\sigma}(t))]\right]\right\}$$

If, instead of $\boldsymbol{\sigma}(t)$, the probability distribution $p_t(\boldsymbol{\sigma})$ is given, this expression is to be averaged over the possible states at time t, with the result:

$$p_{t+1}(\boldsymbol{\sigma})=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left[1+\sigma_i\tanh[\beta h_i(\boldsymbol{\sigma})]\right]p_t(\boldsymbol{\sigma})+\frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\left[1+\sigma_i\tanh[\beta h_i(F_i\boldsymbol{\sigma})]\right]p_t(F_i\boldsymbol{\sigma}) \tag{1.6}$$

in which F_i is the i-th spin-flip operator:

$$F_i\Phi(\boldsymbol{\sigma})\equiv\Phi(\sigma_1,\ldots,\sigma_{i-1},-\sigma_i,\sigma_{i+1},\ldots,\sigma_N)$$

The result (1.6) can also be written in the more familiar form:

$$p_{t+1}(\boldsymbol{\sigma})=\sum_{\boldsymbol{\sigma}'}W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]p_t(\boldsymbol{\sigma}') \tag{1.7}$$

$$W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\equiv\delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'}+\frac{1}{N}\sum_{i=1}^{N}\left\{w_i(F_i\boldsymbol{\sigma})\delta_{\boldsymbol{\sigma},F_i\boldsymbol{\sigma}'}-w_i(\boldsymbol{\sigma})\delta_{\boldsymbol{\sigma},\boldsymbol{\sigma}'}\right\} \tag{1.8}$$

with

$$w_i(\boldsymbol{\sigma})\equiv\frac{1}{2}\left[1-\sigma_i\tanh\left[\beta h_i(\boldsymbol{\sigma})\right]\right] \tag{1.9}$$

Equation (1.7) is the Markov equation corresponding to the sequential process (1.5).

From Discrete to Continuous Times. The formal method to go from any discrete-time Markov process of the form

$$\hat{p}_{n+1}(\boldsymbol{\sigma})=\sum_{\boldsymbol{\sigma}'}W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\hat{p}_n(\boldsymbol{\sigma}') \tag{1.10}$$

to a continuous-time equation, is to assume in addition that the *duration* of each of the above discrete (sequential) iteration steps is a continuous random number. The statistics of these random durations are contained in the function $\pi_m(t)$, which is defined as the probability that at time t precisely m iteration steps have been made. Our new Markov process will now be described by

$$p_t(\boldsymbol{\sigma}) \equiv \sum_{m\geq 0} \pi_m(t)\hat{p}_m(\boldsymbol{\sigma}) = \sum_{m\geq 0} \pi_m(t) \sum_{\boldsymbol{\sigma}'} W^m\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_0(\boldsymbol{\sigma}')$$

and time has become a continuous variable. For $\pi_m(t)$ we make the choice

$$\pi_m(t) \equiv \frac{1}{m!}\left(\frac{t}{\tau}\right)^m e^{-t/\tau} \tag{1.11}$$

with the properties

$$\frac{d}{dt}\pi_{m>0}(t) = \frac{1}{\tau}\left[\pi_{m-1}(t) - \pi_m(t)\right] \qquad \frac{d}{dt}\pi_0(t) = -\frac{1}{\tau}\pi_m(t)$$

From $\langle m\rangle_\pi = t/\tau$ it follows that τ is the average duration of a single discrete iteration step. The above choice for $\pi_m(t)$ allows us to write for the temporal derivative of $p_t(\boldsymbol{\sigma})$:

$$\begin{aligned}\tau\frac{d}{dt}p_t(\boldsymbol{\sigma}) &= \sum_{m>0} \pi_{m-1}(t)\sum_{\boldsymbol{\sigma}'} W^m\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_0(\boldsymbol{\sigma}') - \sum_{m\geq 0} \pi_m(t)\sum_{\boldsymbol{\sigma}'} W^m\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_0(\boldsymbol{\sigma}') \\ &= -p_t(\boldsymbol{\sigma}) + \sum_{\boldsymbol{\sigma}'} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_t(\boldsymbol{\sigma}')\end{aligned}$$

which has the form of a master equation.

This procedure can immediately be applied to equations (1.3,1.7). If for the sequential case in particular we also choose $\tau = \frac{1}{N}$ then in one unit of time each spin will on average have been updated once, and the master equation corresponding to (1.7) acquires the familiar form

$$\frac{d}{dt}p_t(\boldsymbol{\sigma}) = \sum_{i=1}^{N}\left\{w_i(F_i\boldsymbol{\sigma})p_t(F_i\boldsymbol{\sigma}) - w_i(\boldsymbol{\sigma})p_t(\boldsymbol{\sigma})\right\} \tag{1.12}$$

In this equation the quantities $w_i(\boldsymbol{\sigma})$, defined in (1.9), have come to play the role of *transition rates*. If, instead of the specific rule (1.11) we just choose each sequential iteration step to have duration $\tau = 1/N$, one can prove that equation (1.12) will describe the system state for $t \gg \frac{1}{N}$ and $N \to \infty$.

1.2 Interaction Symmetry and Detailed Balance

In the previous section we have not made use of the specific form of the local alignment fields $h_i(\boldsymbol{\sigma})$ (1.2), the relations derived apply equally well to e.g. systems like probabilistic cellular automata. We will now concentrate on properties which depend strongly on the linear dependence of the local alignment fields on the spins $\boldsymbol{\sigma}$.

Symmetric Systems at Zero Temperature. In the deterministic limit $\beta \to \infty$ the stochastic rules (1.1) for parallel dynamics reduce to the deterministic map

$$\sigma_i(t+1) \equiv \text{sgn}\left[h_i(\boldsymbol{\sigma}(t))\right] \tag{1.13}$$

For systems with symmetric interactions, $J_{ij}=J_{ji}$ for all (ij), and stationary external fields, $\theta_i(t)=\theta_i$, we can now construct a Liapunov function (i.e. a function of the dynamic state variables which decreases monotonically and is bounded from below):

$$\tilde{L}(t)\equiv-\sum_{i=1}^{N}|h_i(\boldsymbol{\sigma}(t))|-\sum_{i=1}^{N}\sigma_i(t)\theta_i \tag{1.14}$$

During iteration of the map (1.13) the quantity $\tilde{L}(t)$ evolves according to:

$$\begin{aligned}\tilde{L}(t+1)-\tilde{L}(t)&=-\sum_{i=1}^{N}|h_i(\boldsymbol{\sigma}(t{+}1))|+\sum_{i=1}^{N}\sigma_i(t{+}1)\left[\sum_{j=1}^{N}J_{ij}\sigma_j(t)+\theta_i\right]-\sum_{i=1}^{N}\theta_i\left[\sigma_i(t{+}1)-\sigma_i(t)\right]\\&=-\sum_{i=1}^{N}|h_i(\boldsymbol{\sigma}(t{+}1))|+\sum_{i=1}^{N}\sigma_i(t)h_i(\boldsymbol{\sigma}(t{+}1))\\&=-\sum_{i=1}^{N}|h_i(\boldsymbol{\sigma}(t{+}1))|\left[1-\sigma_i(t{+}2)\sigma_i(t)\right]\;\leq\;0\end{aligned}$$

(where we have used (1.13) and the symmetry of the interaction matrix) So $\tilde{L}(t)$ decreases monotonically until a stage is reached where $\sigma_i(t+2)=\sigma_i(t)$ for all i. From this it follows that parallel systems with symmetric interactions and stationary external fields will in the deterministic limit always end up in a limit cycle with period ≤ 2.

In the deterministic limit $\beta\to\infty$ the stochastic rules (1.5) for sequential dynamics again reduce to a deterministic map (apart from the choice of site to be updated):

$$\sigma_i(t+1)\equiv\delta_{i,\ell_t}\,\mathrm{sgn}\left[h_i(\boldsymbol{\sigma}(t))\right]+(1-\delta_{i,\ell_t})\sigma_i(t) \tag{1.15}$$

For systems with symmetric interactions, $J_{ij}=J_{ji}$ for all (ij), stationary external fields, $\theta_i(t)=\theta_i$, and without self-interactions, $J_{ii}=0$ for all i, we can again construct a Liapunov function:

$$L(t)\equiv-\frac{1}{2}\sum_{ij=1}^{N}\sigma_i(t)J_{ij}\sigma_j(t)-\sum_{i=1}^{N}\sigma_i(t)\theta_i \tag{1.16}$$

Clearly this quantity is bounded from below. If we put $\ell_t\equiv\ell$ then $L(t)$ evolves during iteration of the map (1.15) according to:

$$\begin{aligned}L(t+1)-L(t)&=\left[\sigma_\ell(t)-\sigma_\ell(t{+}1)\right]\left\{\frac{1}{2}\sum_{j=1}^{N}J_{\ell j}\sigma_j(t)+\frac{1}{2}\sum_{i=1}^{N}J_{i\ell}\sigma_i(t)+\theta_\ell\right\}\\&=\left[\sigma_\ell(t)-\sigma_\ell(t{+}1)\right]h_\ell(\boldsymbol{\sigma}(t))\\&=-|h_\ell(\boldsymbol{\sigma}(t))|\left[1-\sigma_\ell(t)\sigma_\ell(t{+}1)\right]\;\leq\;0\end{aligned}$$

(where we have used (1.15), symmetry of the interaction matrix and the absence of diagonal interactions J_{ii}) So $L(t)$ decreases monotonically until a stage is reached where $\sigma_i(t{+}1)=\sigma_i(t)$ for all i. We conclude that sequential systems with symmetric interactions without diagonal terms ($J_{ii} = 0$ for all i) and stationary external fields will in the deterministic limit always end up in a stationary state.

Detailed Balance. The results obtained above suggest that symmetric systems, where $J_{ij} = J_{ji}$ for all (ij), represent a special class. We now show how interaction symmetry is closely related to the detailed balance property, and derive a number of consequences. A Markov process of the form (1.3,1.7), i.e.

$$p_{t+1}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_t(\boldsymbol{\sigma}') \tag{1.17}$$

$$W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] \in [0,1] \qquad\qquad \sum_{\boldsymbol{\sigma}} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] = 1$$

(where the conditions on the transition matrix ensure a probabilistic interpretation of $p_t(\boldsymbol{\sigma})$) is said to obey detailed balance if there exists a stationary solution $p_\infty(\boldsymbol{\sigma})$ of (1.17) with the property:

$$W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_\infty(\boldsymbol{\sigma}') = W\left[\boldsymbol{\sigma}';\boldsymbol{\sigma}\right] p_\infty(\boldsymbol{\sigma}) \qquad \text{for all } \boldsymbol{\sigma},\boldsymbol{\sigma}' \tag{1.18}$$

For a stationary distribution to exist we have to require external fields to be stationary: $\theta_i(t) = \theta_i$. All distributions $p_\infty(\boldsymbol{\sigma})$ which satisfy (1.18) are stationary solutions of (1.17), as can be easily verified by substitution. The converse is not true. Detailed balance is a special feature which usually simplifies the calculation of the stationary probability distribution $p_\infty(\boldsymbol{\sigma})$. It states that, in addition to the probability distribution being stationary, it describes *equilibrium* in the sense that there is no net probability current between any two microscopic system states. It is not a necessary condition for the existence or uniqueness of (or the approach to) a stationary solution of (1.17). These latter issues will be addressed in a following section.

Parallel Dynamics. For parallel dynamics the transition matrix is given by (1.4) and the detailed balance condition (1.18) becomes

$$\frac{e^{\beta\sum_{i=1}^N \sigma_i h_i(\boldsymbol{\sigma}')} p_\infty(\boldsymbol{\sigma}')}{\prod_{i=1}^N \cosh[\beta h_i(\boldsymbol{\sigma}')]} = \frac{e^{\beta\sum_{i=1}^N \sigma_i' h_i(\boldsymbol{\sigma})} p_\infty(\boldsymbol{\sigma})}{\prod_{i=1}^N \cosh[\beta h_i(\boldsymbol{\sigma})]} \qquad \text{for all } \boldsymbol{\sigma},\boldsymbol{\sigma}' \tag{1.19}$$

The transition matrix (1.4) describes an *ergodic* system, i.e. from any initial state $\boldsymbol{\sigma}$ one can reach any final state $\boldsymbol{\sigma}'$ with nonzero probability in a finite number of steps (in this case: one). As a consequence all stationary probabilities $p_\infty(\boldsymbol{\sigma})$ must be non-zero. Without loss of generality we can therefore put:

$$p_\infty(\boldsymbol{\sigma}) \equiv e^{\beta\left[\sum_{i=1}^N \theta_i\sigma_i + K(\boldsymbol{\sigma})\right]} \prod_{i=1}^N \cosh[\beta h_i(\boldsymbol{\sigma})] \tag{1.20}$$

which, in combination with the definition (1.2) of the local fields $h_i(\boldsymbol{\sigma})$, simplifies the detailed balance condition to:

$$K(\boldsymbol{\sigma}) - K(\boldsymbol{\sigma}') = \sum_{ij=1}^N \sigma_i \left[J_{ij} - J_{ji}\right] \sigma_j' \qquad \text{for all } \boldsymbol{\sigma},\boldsymbol{\sigma}' \tag{1.21}$$

If we now take the average of the above expression over the 2^N states $\boldsymbol{\sigma}'$ we obtain:

$$K(\boldsymbol{\sigma}) = \langle K(\boldsymbol{\sigma}')\rangle_{\boldsymbol{\sigma}'} \qquad \text{for all } \boldsymbol{\sigma}$$

This implies that K can only be a constant, which is determined by normalising (1.19). Therefore, if detailed balance holds, then the corresponding equilibrium distribution must be:

$$p_{\rm eq}(\boldsymbol{\sigma}) \sim e^{\beta\sum_{i=1}^N \theta_i\sigma_i}\prod_{i=1}^N \cosh[\beta h_i(\boldsymbol{\sigma})] \tag{1.22}$$

For symmetric systems detailed balance indeed holds and (1.22) solves (1.19), since $K(\boldsymbol{\sigma}) =$ constant solves the reduced problem (1.21). For non-symmetric systems, however, there can be no equilibrium, since for $K(\boldsymbol{\sigma}) =$ constant it follows from (1.21) that there is only going to be detailed balance if the interaction matrix has the property

$$\sum_{ij=1}^N \sigma_i\,[J_{ij}-J_{ji}]\,\sigma_j' = 0 \qquad \text{for all } \boldsymbol{\sigma},\boldsymbol{\sigma}'$$

For $2^{N-1} \geq N$ (or: $N \geq 2$) the vector pairs $(\boldsymbol{\sigma},\boldsymbol{\sigma}')$ span the space of all $N\times N$ matrices and no such non-symmetric interaction matrices will exist, whereas for $N=1$ there is simply no non-symmetric interaction matrix. For parallel dynamics interaction symmetry implies detailed balance and vice versa.

Sequential Dynamics Without Self-Interactions. For sequential dynamics the transition matrix is given by (1.8) and the detailed balance condition (1.18) simplifies to

$$\frac{e^{\beta\sigma_i h_i(F_i\boldsymbol{\sigma})}p_\infty(F_i\boldsymbol{\sigma})}{\cosh\left[\beta h_i(F_i\boldsymbol{\sigma})\right]} = \frac{e^{-\beta\sigma_i h_i(\boldsymbol{\sigma})}p_\infty(\boldsymbol{\sigma})}{\cosh\left[\beta h_i(\boldsymbol{\sigma})\right]} \qquad \text{for all } \boldsymbol{\sigma} \text{ and all } i$$

The sequential transition matrix (1.8) also describes an *ergodic* system; here we can get from any initial state $\boldsymbol{\sigma}$ to any desired final state $\boldsymbol{\sigma}'$ with non-zero probability in at most N iterations. Since all stationary probabilities $p_\infty(\boldsymbol{\sigma})$ must therefore be non-zero, we can write:

$$p_\infty(\boldsymbol{\sigma}) \equiv e^{\beta\left[\sum_{i=1}^N \theta_i\sigma_i + \frac{1}{2}\sum_{i\neq j}\sigma_i J_{ij}\sigma_j + K(\boldsymbol{\sigma})\right]} \tag{1.23}$$

Using relations like

$$\sum_{k\neq l} J_{kl}F_i(\sigma_k\sigma_l) = \sum_{k\neq l} J_{kl}\sigma_k\sigma_l - 2\sigma_i\sum_{k\neq i}\left[J_{ik}+J_{ki}\right]\sigma_k$$

we can now simplify the detailed balance condition to:

$$K(F_i\boldsymbol{\sigma}) - K(\boldsymbol{\sigma}) = \sigma_i\sum_{k\neq i}\left[J_{ik}-J_{ki}\right]\sigma_k \qquad \text{for all } \boldsymbol{\sigma} \text{ and all } i \tag{1.24}$$

If to this expression we apply the general identity

$$\left[1-F_i\right]f(\boldsymbol{\sigma}) = 2\sigma_i\langle\sigma_i f(\boldsymbol{\sigma})\rangle_{\sigma_i}$$

we find for $i \neq j$:

$$[F_j - 1][F_i - 1]K(\boldsymbol{\sigma}) = -2\sigma_i\sigma_j\,[J_{ij} - J_{ji}] \qquad \text{for all } \boldsymbol{\sigma} \text{ and all } i \neq j$$

The left-hand side is symmetric under permutation of the pair (i, j), which immediately implies that the interaction matrix must be symmetric: $J_{ij} = J_{ji}$ for all (i, j). We find the trivial solution $K(\boldsymbol{\sigma})$ = constant, detailed balance therefore holds and the corresponding equilibrium distribution is

$$p_\infty(\boldsymbol{\sigma}) \ \sim \ e^{-\beta H(\boldsymbol{\sigma})} \qquad H(\boldsymbol{\sigma}) \equiv -\frac{1}{2}\sum_{ij=1}^{N}\sigma_i J_{ij}\sigma_j - \sum_{i=1}^{N}\theta_i\sigma_i \tag{1.25}$$

We may conclude that for sequential systems without self-interactions interaction symmetry implies detailed balance and *vice versa.*

This result establishes the link with conventional equilibrium statistical mechanics. The noise parameter β can formally be identified with the inverse ‘temperature’ in equilibrium, $\beta = T^{-1}$, and the function $H(\boldsymbol{\sigma})$ is the usual Ising spin Hamiltonian. Indeed for symmetric systems without self-interactions the sequential stochastic process (1.5) simply reduces to a Glauber (1963) dynamics associated with the Hamiltonian H, since the transition probabilities $w_i(\boldsymbol{\sigma})$ (1.9) for the state change $\boldsymbol{\sigma} \to F_i\boldsymbol{\sigma}$ can be expressed in terms of the resulting change in energy $\Delta_i H(\boldsymbol{\sigma})$ as

$$w_i(\boldsymbol{\sigma}) = \left[1 + e^{\beta\Delta_i H(\boldsymbol{\sigma})}\right]^{-1} \qquad \Delta_i H(\boldsymbol{\sigma}) \equiv H(F_i\boldsymbol{\sigma}) - H(\boldsymbol{\sigma})$$

Sequential Dynamics with Self-Interactions. In this final case the situation is more complicated. In principle detailed balance may hold for symmetric and non-symmetric systems. However, one can show that such cases must be pathological ones, since only for very specific choices for the matrix elements $\{J_{ij}\}$ and low-dimensional systems can the corresponding requirements be met (like systems with self-interactions *only*).

1.3 Equilibrium Statistical Mechanics

The Free Energy. For systems with symmetric interaction matrices (in the sequential case: without self-interactions) and stationary external fields detailed balance holds and we know the equilibrium probability distribution. For sequential dynamics this distribution has the Boltzmann form and we can apply standard equilibrium statistical mechanics. In particular we can define the partition function Z and the free energy F:

$$p_{\text{eq}}(\boldsymbol{\sigma}) \ \sim \ e^{-\beta H(\boldsymbol{\sigma})} \qquad H(\boldsymbol{\sigma}) \equiv -\frac{1}{2}\sum_{ij=1}^{N}\sigma_i J_{ij}\sigma_j - \sum_{i=1}^{N}\theta_i\sigma_i \tag{1.26}$$

$$Z \equiv \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma})} \qquad F \equiv -\frac{1}{\beta}\log Z$$

These can be used in the usual manner to generate equilibrium averages $\langle\ldots\rangle_{\rm eq}$ of state variables. Taking derivatives with respect to external fields θ_i and interactions J_{ij} produces

$$-\frac{\partial F}{\partial\theta_i}=\langle\sigma_i\rangle_{\rm eq}\qquad\qquad -\frac{\partial F}{\partial J_{ij}}=\langle\sigma_i\sigma_j\rangle_{\rm eq}$$

whereas the equilibrium average of any arbitrary state variable $f(\boldsymbol{\sigma})$ can be obtained by adding suitable generating terms to the Hamiltonian:

$$H(\boldsymbol{\sigma})\to H(\boldsymbol{\sigma})+\lambda f(\boldsymbol{\sigma})\qquad\qquad \langle f\rangle_{\rm eq}=\lim_{\lambda\to 0}\frac{\partial F}{\partial\lambda}$$

In the parallel case we can again formally write the equilibrium probability distribution in the Boltzmann form and define a corresponding partition function $\tilde{Z}$ and a free energy $\tilde{F}$:

$$p_{\rm eq}(\boldsymbol{\sigma})\;\sim\;e^{-\beta\tilde{H}(\boldsymbol{\sigma})}\qquad \tilde{H}(\boldsymbol{\sigma})\equiv-\sum_{i=1}^N\theta_i\sigma_i-\frac{1}{\beta}\sum_{i=1}^N\log 2\cosh[\beta h_i(\boldsymbol{\sigma})]\tag{1.27}$$

$$\tilde{Z}\equiv\sum_{\boldsymbol{\sigma}}e^{-\beta\tilde{H}(\boldsymbol{\sigma})}\qquad\qquad \tilde{F}\equiv-\frac{1}{\beta}\log\tilde{Z}$$

with which we can obtain equilibrium averages:

$$\tilde{H}(\boldsymbol{\sigma})\to\tilde{H}(\boldsymbol{\sigma})+\lambda f(\boldsymbol{\sigma})\qquad\qquad \langle f\rangle_{\rm eq}=\lim_{\lambda\to 0}\frac{\partial\tilde{F}}{\partial\lambda}$$

But we have to keep in mind that thermodynamic relations involving derivations with respect to β will no longer be valid, and that by taking derivatives with respect to external fields or interactions different equilibrium averages will be generated. For instance

$$-\frac{\partial\tilde{F}}{\partial\theta_i}=\langle\sigma_i\rangle_{\rm eq}+\langle\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle_{\rm eq}$$

$$i\neq j:\qquad -\frac{\partial\tilde{F}}{\partial J_{ij}}=\langle\sigma_i\tanh[\beta h_j(\boldsymbol{\sigma})]\rangle_{\rm eq}+\langle\sigma_j\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle_{\rm eq}$$

$$i=j:\qquad -\frac{\partial\tilde{F}}{\partial J_{ii}}=\langle\sigma_i\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle_{\rm eq}$$

One can use $\langle\sigma_i\rangle_{\rm eq}=\langle\tanh[\beta h_i(\boldsymbol{\sigma})]\rangle_{\rm eq}$, which can derived directly from the equilibrium equation $p_{t+1}(\boldsymbol{\sigma})=p_t(\boldsymbol{\sigma})$, to simplify the first of these identities.

By writing the term $\prod_i 2\cosh[\beta h_i(\boldsymbol{\sigma})]$ as a trace over an auxiliary set of spin variables $\{\sigma_i\}$ we can write the parallel dynamics partition function $\tilde{Z}$ in a form which is more reminiscent of the sequential dynamics one, is more suitable for averaging over quenched disorder and also hints at a simple relation between the free energies F and $\tilde{F}$:

$$\tilde{Z}=\sum_{\boldsymbol{\sigma}}\sum_{\boldsymbol{\sigma}'}e^{-\beta\left[-\frac{1}{2}\sum_{ij=1}^N\sigma_iJ_{ij}\sigma'_j-\frac{1}{2}\sum_{ij=1}^N\sigma'_iJ_{ij}\sigma_j-\sum_{i=1}^N\theta_i\sigma_i-\sum_{i=1}^N\theta_i\sigma'_i\right]}\tag{1.28}$$

Note, finally, that from the equilibrium distributions we can recover in the deterministic limit the Liapunov functions (1.14) and (1.16):

$$L(\boldsymbol{\sigma}) = H(\boldsymbol{\sigma}) \qquad \tilde{L}(\boldsymbol{\sigma}) = \lim_{\beta\to\infty} \tilde{H}(\boldsymbol{\sigma})$$

Relation Between Free Energies: a Mean Field Example. We now turn to a simple toy model to illustrate similarities and differences between sequential and parallel dynamics from the point of view of equilibrium statistical mechanical calculations. We will choose a system with uniform infinite-range pair interactions and zero external fields:

$$J_{ij} = J_{ji} \equiv \frac{J}{N} \quad (i \neq j), \qquad \theta_i = 0$$

For which we obtain:

$$\lim_{N\to\infty} \frac{F}{N} = -\lim_{N\to\infty} \frac{1}{\beta N} \log \sum_{\boldsymbol{\sigma}} e^{\beta N[\frac{1}{2}Jm^2(\boldsymbol{\sigma})]}$$

$$\lim_{N\to\infty} \frac{\tilde{F}}{N} = -\lim_{N\to\infty} \frac{1}{\beta N} \log \sum_{\boldsymbol{\sigma}} e^{N[\log 2\cosh[\beta Jm(\boldsymbol{\sigma})]]}$$

with the average activity or magnetisation $m(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{k=1}^{N}\sigma_k$. Apparently we only have to count the number of microscopic states with a prescribed magnetisation $m = 2n/N - 1$ (where n is the number of sites with $\sigma = 1$), in expressions of the form

$$\begin{aligned} \frac{1}{N}\log\sum_{\boldsymbol{\sigma}} e^{NU[m(\boldsymbol{\sigma})]} &= \frac{1}{N}\log\sum_{n=0}^{N}\binom{N}{n} e^{NU[2n/N-1]} \\ &= \frac{1}{N}\log\int_{-1}^{1} dm\; e^{N[\log 2 - c^*(m) + U[m]]} \\ &= \log 2 + \max_m \{U[m] - c^*(m)\} \end{aligned}$$

with

$$c^*(m) \equiv \frac{1}{2}(1+m)\log(1+m) + \frac{1}{2}(1-m)\log(1-m)$$

We used Stirling's formula to obtain the leading term of the factorials (only contributions which are exponential in N will survive the limit $N \to \infty$), which led us to a saddle-point problem. The result can readily be applied to obtain the free energies:

$$\beta \lim_{N\to\infty} \frac{F}{N} = -\log 2 + \min_m \left\{c^*(m) - \frac{1}{2}\beta Jm^2\right\} \tag{1.29}$$

$$\beta \lim_{N\to\infty} \frac{\tilde{F}}{N} = -2\log 2 + \min_m \{c^*(m) - \log\cosh[\beta Jm]\} \tag{1.30}$$

The shapes of the functions to be minimised are shown in figure 1.1. The equations from which to solve the suprema are easily obtained by differentiation, since $\frac{d}{dm}c^*(m) = \tanh^{-1}(m)$. For sequential dynamics we find the Curie-Weiss law:

$$m = \tanh[\beta Jm] \tag{1.31}$$

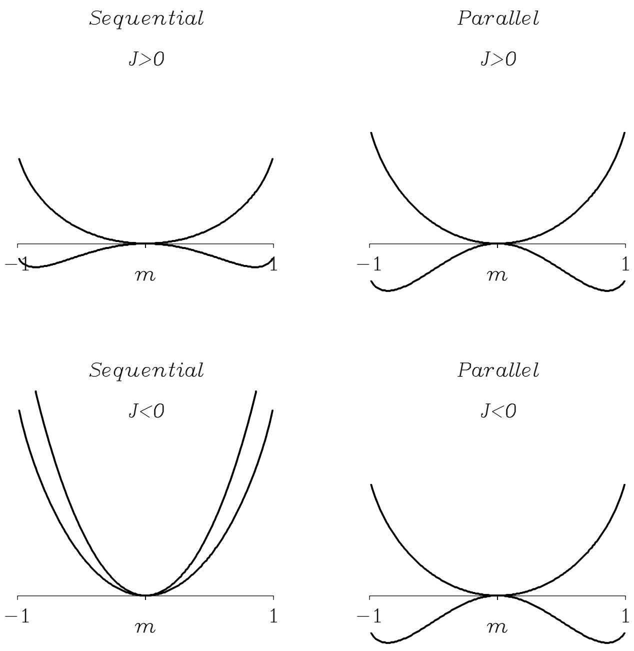


Figure 1.1: The minimisation problems determining the free energies for sequential (left) and parallel dynamics (right). In each graph the upper line refers to the case $\beta|J| < 1$ and the lower line to the case $\beta|J| > 1$.

For parallel dynamics:

$$m = \tanh\left[\beta J \tanh[\beta J m]\right] \tag{1.32}$$

The solutions of the latter equation (1.32) again obey a Curie-Weiss law, which follows from the following argument. The definition $\hat{m} \equiv \tanh[\beta|J|m]$ transforms (1.32) into

$$m = \tanh[\beta|J|\hat{m}] \qquad \hat{m} = \tanh[\beta|J|m]$$

from which one derives

$$0 \leq [m - \hat{m}]^2 = [m - \hat{m}]\left[\tanh[\beta|J|\hat{m}] - \tanh[\beta|J|m]\right] \leq 0$$

This immediately implies:

$$m = \tanh[\beta|J|m] \tag{1.33}$$

If $J \geq 0$ the two types of dynamics lead to the same thermodynamics. At temperatures above $T_c \equiv J$, both minimisation problems are solved by $m = 0$ (see figure 1.1), describing a paramagnetic state. Below T_c they are solved by the two non-zero solutions of the Curie-Weiss law (1.31). Furthermore, using the identity $c^*(\tanh x) = x \tanh x - \log\cosh x$, we obtain from (1.29,1.30) the relation $\lim_{N\to\infty} \tilde{F}/N = 2\lim_{N\to\infty} F/N$. If $J < 0$, however, the two types of dynamics give quite different results. For sequential dynamics the required minimum is obtained for $m = 0$ (the paramagnetic state). For parallel dynamics the minimisation problem is invariant under $J \to -J$; therefore the thermodynamic behaviour is again of the

Curie-Weiss type (as indicated by figure 1.1 and equation (1.33)), with an ordered state for temperatures below $T_c \equiv |J|$.

The explanation of the difference between the two types of dynamics for $J < 0$ is clear from studying the evolution in time of the order parameter m. As we will see in a subsequent chapter, for the present (toy) model in the limit $N \to \infty$ the magnetisation evolves in time according to the deterministic mean-field laws

$$\frac{d}{dt}m = \tanh[\beta J m] - m \qquad\qquad m(t+1) = \tanh[\beta J m(t)]$$

for sequential and parallel dynamics, respectively. For $J < 0$ the sequential system will always decay towards the paramagnetic state $m = 0$, whereas for sufficiently large β the parallel system can enter into the stable limit-cycle $m(t) = M_\beta(-1)^t$ (where M_β is the non-zero solution of (1.33)). The concepts of 'distance' and 'local minima' are quite different for the two types of dynamics; in contrast to the sequential case, parallel dynamics allows the system to make the transition $m \to -m$ in thermal equilibrium.

1.4 The Approach to a Stationary Distribution

In the previous sections we have been concerned with properties of our system in the case where a steady state is reached (i.e. where the system is described by a probability distribution $p_\infty(\boldsymbol{\sigma})$ which is stationary). In this last section of the chapter on microscopic dynamics we turn to the important questions of (i) in which cases the system can indeed be guaranteed to go to such a state, and (ii) whether this state is unique. There are many approaches to this problem, most of which, however, are not sufficiently general for our purpose since they are based on the detailed balance property (which in our case would exclude all non-symmetric networks). It will turn out that all finite systems of the classes (1.3,1.7) considered so far (sequential and parallel, with and without symmetric interactions or self-interactions) can be guaranteed to converge towards a unique stationary probability distribution. The only relevant property is that both types of transition matrices considered, (1.3) and (1.7), are ergodic: whatever the initial conditions $p_0(\boldsymbol{\sigma})$, there exists a time T such that

$$p_t(\boldsymbol{\sigma}) > 0 \qquad \text{for all } \boldsymbol{\sigma} \text{ and all } t \geq T \tag{1.34}$$

For parallel dynamics (1.3) just one iteration is needed for obtaining positive probabilities only: $T_{\text{par}} = 1$. For sequential dynamics (1.7) at most N iterations are needed: $T_{\text{seq}} = N$.

Existence of a Stationary Probability Distribution. We consider Markov processes of the type (1.17). Conservation of probability guarantees that the transition matrix has a left eigenvector with eigenvalue one:

$$\sum_{\boldsymbol{\sigma}} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] = 1$$

For any finite-dimensional matrix the spectrum of left eigenvalues is identical to the spectrum of right eigenvalues, since

$$|\det\left[W - \lambda I\right]| = |\det\left[W - \lambda I\right]^\dagger|$$

Therefore there exists at least one non-trivial right eigenvector ϕ with eigenvalue one:

$$\sum_{\boldsymbol{\sigma}'} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\phi(\boldsymbol{\sigma}') = \phi(\boldsymbol{\sigma}) \tag{1.35}$$

We will now prove that for ergodic systems (1.34) there exists an eigenvector of the type (1.35) with positive components only. We define S as the set of all system states $\boldsymbol{\sigma}$ for which $\phi(\boldsymbol{\sigma}) > 0$. From (1.35) we can derive

$$\sum_{\boldsymbol{\sigma}\in S}\phi(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}\in S}\left[\sum_{\boldsymbol{\sigma}'\in S} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\phi(\boldsymbol{\sigma}') + \sum_{\boldsymbol{\sigma}'\notin S} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\phi(\boldsymbol{\sigma}')\right]$$

or

$$\sum_{\boldsymbol{\sigma}'\in S}\left[1-\sum_{\boldsymbol{\sigma}\in S} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\right]\phi(\boldsymbol{\sigma}') = \sum_{\boldsymbol{\sigma}\in S}\sum_{\boldsymbol{\sigma}'\notin S} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\phi(\boldsymbol{\sigma}')$$

By construction the left-hand side of this identity is non-negative and the right-hand side is non-positive, both must therefore be zero:

$$\text{for all } \boldsymbol{\sigma}'\in S,\ \boldsymbol{\sigma}\notin S: \qquad W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] = 0$$

$$\text{for all } \boldsymbol{\sigma}\in S,\ \boldsymbol{\sigma}'\notin S: \qquad W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] = 0$$

The only allowed microscopic transitions $\boldsymbol{\sigma}\to\boldsymbol{\sigma}'$ are apparently of the type

$$\begin{array}{ccc}\boldsymbol{\sigma}\in S & \to & \boldsymbol{\sigma}'\in S\\ \boldsymbol{\sigma}\notin S & \to & \boldsymbol{\sigma}'\notin S\end{array}$$

Suppose S is not empty. Our systems being ergodic (1.34) now immediately implies that S must contain *all* states. In other words: if ϕ has positive components, then *all* components must be positive. In the same way we could have defined S to be the set of states $\boldsymbol{\sigma}$ with $\phi(\boldsymbol{\sigma}) < 0$, in which case the derivation would have led to the result: if ϕ has negative components, then *all* components must be negative. Since from any solution ϕ of (1.35) with negative components only we can obtain one with positive components only by switching $\phi\to-\phi$ (which can subsequently be properly normalised), we have proven the existence of a stationary probability distribution.

Convergence and Uniqueness. We now turn to the evolution in time of the difference ψ_t between two individual probability distributions p_t^a and p_t^b, which are obtained as a result of iterating (1.17) from different initial conditions a and b. Due to the linearity of the problem ψ_t is itself again a solution of (1.17):

$$\psi_t(\boldsymbol{\sigma}) \equiv p_t^a(\boldsymbol{\sigma}) - p_t^b(\boldsymbol{\sigma}) \qquad\qquad \psi_{t+1}(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}'} W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]\psi_t(\boldsymbol{\sigma})$$

Since the process (1.17) conserves probability, $\psi_t(\boldsymbol{\sigma})$ has the additional property

$$\sum_{\boldsymbol{\sigma}}\psi_t(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\sigma}}\left[p_t^a(\boldsymbol{\sigma}) - p_t^b(\boldsymbol{\sigma})\right] = 0$$

The proof of convergence is based on the existence of a Liapunov function $U(t)$. First we divide the set of all system states $\boldsymbol{\sigma}$ into subsets, according to the sign of the corresponding component of ψ_t:

$$\begin{aligned} S_+(t) &\equiv \text{all } \boldsymbol{\sigma} \text{ with } \psi_t(\boldsymbol{\sigma}) > 0 \\ S_0(t) &\equiv \text{all } \boldsymbol{\sigma} \text{ with } \psi_t(\boldsymbol{\sigma}) = 0 \\ S_-(t) &\equiv \text{all } \boldsymbol{\sigma} \text{ with } \psi_t(\boldsymbol{\sigma}) < 0 \end{aligned}$$

We now define $U(t)$ as the sum of all positive components of the vector ψ_t:

$$U(t) \equiv \sum_{\boldsymbol{\sigma}\in S_+(t)} \psi_t(\boldsymbol{\sigma}) \tag{1.36}$$

Clearly $U(t) \geq 0$. Furthermore the sum $U(t)$ can be shown to decrease monotonically with time:

$$\begin{aligned} U(t+1) - U(t) &= \sum_{\boldsymbol{\sigma}\in S_+(t+1)} \psi_{t+1}(\boldsymbol{\sigma}) - \sum_{\boldsymbol{\sigma}\in S_+(t)} \psi_t(\boldsymbol{\sigma}) \\ &= - \sum_{\boldsymbol{\sigma}'\in S_+(t)} \psi_t(\boldsymbol{\sigma}') \left[1 - \sum_{\boldsymbol{\sigma}\in S_+(t+1)} W[\boldsymbol{\sigma};\boldsymbol{\sigma}']\right] + \sum_{\boldsymbol{\sigma}'\in S_-(t)} \psi_t(\boldsymbol{\sigma}') \sum_{\boldsymbol{\sigma}\in S_+(t+1)} W[\boldsymbol{\sigma};\boldsymbol{\sigma}'] \;\leq\; 0 \end{aligned} \tag{1.37}$$

We are led to the conclusion that $U(t)$ must tend to a limit: $\lim_{t\to\infty} U(t) = U(\infty) \geq 0$ exists. Since the sum of all components of ψ_t is zero, the sum over all negative components of ψ_t must tend to the limit $-U(\infty)$. According to (1.37), $U(t)$ being stationary implies the following:

$$\begin{aligned} &\text{for all } \boldsymbol{\sigma}' \in S_+(t),\ \boldsymbol{\sigma} \notin S_+(t+1): \quad W[\boldsymbol{\sigma};\boldsymbol{\sigma}'] = 0 \\ &\text{for all } \boldsymbol{\sigma}' \in S_-(t),\ \boldsymbol{\sigma} \in S_+(t+1): \quad W[\boldsymbol{\sigma};\boldsymbol{\sigma}'] = 0 \end{aligned}$$

If we now run through the same analysis with ψ_t being replaced by $-\psi_t$, we find similar relations in which the sets S_+ and S_- have changed places. These relations can be combined with the ones above to give a quite restrictive set of conditions on the transition matrix itself:

$$\begin{aligned} &\text{for all } \boldsymbol{\sigma}' \in S_+(t),\ \boldsymbol{\sigma} \notin S_+(t+1): \quad W[\boldsymbol{\sigma};\boldsymbol{\sigma}'] = 0 \\ &\text{for all } \boldsymbol{\sigma}' \in S_-(t),\ \boldsymbol{\sigma} \notin S_-(t+1): \quad W[\boldsymbol{\sigma};\boldsymbol{\sigma}'] = 0 \end{aligned}$$

The only allowed microscopic transitions $\boldsymbol{\sigma} \to \boldsymbol{\sigma}'$ are apparently of the type

$$\begin{aligned} \boldsymbol{\sigma} \in S_+(t) &\;\to\; \boldsymbol{\sigma}' \in S_+(t+1) \\ \boldsymbol{\sigma} \in S_0(t) &\;\to\; \boldsymbol{\sigma}' \in S_+(t+1) \bigcup S_0(t+1) \bigcup S_-(t+1) \\ \boldsymbol{\sigma} \in S_-(t) &\;\to\; \boldsymbol{\sigma}' \in S_-(t+1) \end{aligned}$$

Suppose $S_+(t)$ is not empty. If we now were to prepare our system in a microscopic configuration $\boldsymbol{\sigma}_0 \in S_+(t)$, then at *any* time $t' > t$ we would only be able to find the system in a state in the set $S_+(t')$. Our systems being ergodic (1.34) now immediately implies that both $S_-(t')$ and $S_0(t')$ must be empty for all $t' \geq t + T$. If, on the other hand, $S_-(t)$ is not empty we can prepare our system in a microscopic configuration $\boldsymbol{\sigma}_0 \in S_-(t)$ and find similarly that both $S_+(t')$ and $S_0(t')$ must be empty for all $t' \geq t + T$. These two situations are exclusive.

Therefore we can conclude that for an ergodic system one of the two sets $S_+(t)$ and $S_-(t)$ must be empty. This, in turn, implies that *all components of ψ_t must be zero*, since

$$\text{all } \psi_t(\boldsymbol{\sigma}) \leq 0 \;\; \text{and} \;\; \sum_{\boldsymbol{\sigma}} \psi_t(\boldsymbol{\sigma}) = 0 \quad \rightarrow \quad \psi_t(\boldsymbol{\sigma}) = 0 \text{ for all } \boldsymbol{\sigma}$$

$$\text{all } \psi_t(\boldsymbol{\sigma}) \geq 0 \;\; \text{and} \;\; \sum_{\boldsymbol{\sigma}} \psi_t(\boldsymbol{\sigma}) = 0 \quad \rightarrow \quad \psi_t(\boldsymbol{\sigma}) = 0 \text{ for all } \boldsymbol{\sigma}$$

In terms of our two evolving probability distributions this implies

$$\lim_{t\to\infty} \left[p_t^a(\boldsymbol{\sigma}) - p_t^b(\boldsymbol{\sigma})\right] = 0 \qquad \text{for all } \boldsymbol{\sigma}$$

In particular we can choose one of the two to be the stationary probability distribution $p_\infty(\boldsymbol{\sigma})$, the existence of which we have already demonstrated, which gives:

$$\lim_{t\to\infty} p_t(\boldsymbol{\sigma}) = p_\infty(\boldsymbol{\sigma}) \qquad \text{for all } \boldsymbol{\sigma} \tag{1.38}$$

This completes the proof. Every probability distribution will converge towards p_∞. The existence of more than one stationary probability distribution would lead to a contradiction in (1.38) and is therefore ruled out.

Chapter 2

Recurrent Networks in Equilibrium

In this chapter we will restrict ourselves to symmetric networks which obey detailed balance, so that we know the equilibrium probability distribution and equilibrium statistical mechanics applies. In the case of sequential dynamics we will accordingly not allow for the presence of self-interactions. We will study in detail the Hopfield model, which is the archetypical model to describe the functioning of symmetric neural networks as associative memories. The basic recipe is the following:

- Represent each of the items or patterns to be stored (pictures, words, etc.) as an N-bit vector $\boldsymbol{\xi}^{\mu} \in \{-1,1\}^N$.

- Construct neural interactions $\{J_{ij}\}$ and thresholds $\{\theta_i\}$ such that the microscopic ground state configurations of the network are located at (or very close to) these pattern vectors $\boldsymbol{\xi}^{\mu}$, which are to function as attractors in phase space.

- If now we are given an input to be recognised, we choose this input to be the initial spin configuration $\boldsymbol{\sigma}(0)$ of our system. From this initial state the system is allowed to evolve in time autonomously, which in the zero temperature case will lead to the nearest attractor (in some topological sense).

(see figure 2.1). If our choice of interactions and thresholds was indeed such that only the patterns $\boldsymbol{\xi}^{\mu}$ act as attractors, then the final state reached $\boldsymbol{\sigma}(\infty)$ can be interpreted as the pattern recognized by network from the input $\boldsymbol{\sigma}(0)$. For such a program to work we need spin systems with ergodicity breaking: in the limit $N \to \infty$ the system will on finite time-scales have to be confined to a restricted region of phase space, the location of which is to depend strongly on the initial conditions. Secondly, since the number of attractors, i.e. the number of patterns stored, should be preferably large, we need systems with a large number of ergodic components.

To illustrate some of the fundamental features of the recipe, let us first consider the simplest case and try to store just a single pattern $\boldsymbol{\xi} \in \{-1,1\}$ in a noiseless (zero-temperature) infinite-range network. Obvious candidates for interactions and thresholds would be

$$J_{ij} = \frac{1}{N}\xi_i\xi_j \qquad \theta_i = 0, \tag{2.1}$$

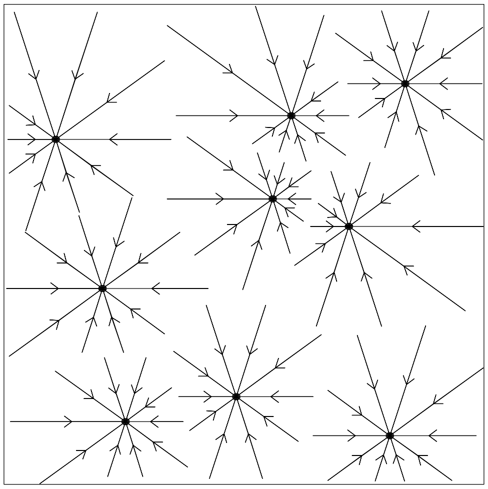


Figure 2.1: Information storage and retrieval in Ising spin neural networks through the creation of attractors in phase space. Patterns $\boldsymbol{\xi}^{\mu}$ to be retrieved: $\bullet$. If the interactions are choosen to be symmetric, as in the Hopfield model, the attractors will have to be fixed-points.

(for sequential dynamics we put $J_{ii} = 0$ for all i). The zero temperature Liapunov functions (1.14,1.16) become:

$$\tilde{L}(t) = -|\sum_{j=1}^{N} \xi_i \sigma_i(t)| \qquad L(t) = \frac{1}{2} - \frac{1}{2N}\left[\sum_{i=1}^{N} \xi_i \sigma_i(t)\right]^2 \tag{2.2}$$

(for parallel and sequential dynamics, respectively), from which we immediately obtain

$$\begin{aligned} \textstyle\sum_{i=1}^{N} \xi_i \sigma_i(t) > 0: &\qquad \boldsymbol{\sigma}(\infty) = \boldsymbol{\xi} \\ \textstyle\sum_{i=1}^{N} \xi_i \sigma_i(t) < 0: &\qquad \boldsymbol{\sigma}(\infty) = -\boldsymbol{\xi} \end{aligned}$$

We could also have performed the gauge transformation $\sigma_i \to \sigma_i \xi_i$, which would have led us to an ordinary infinite-range ferromagnet. The system can indeed reconstruct dynamically the original pattern $\boldsymbol{\xi}$ from an input vector $\boldsymbol{\sigma}(0)$. What we also note, however, is that *en passant* we have created an additional attractor: the state $-\boldsymbol{\xi}$. This property is shared by all models in which external fields θ_i are zero, where the Hamiltonians $H(\boldsymbol{\sigma})$ (for sequential dynamics) and $\tilde{H}(\boldsymbol{\sigma})$ (for parallel dynamics) are invariant under an overall sign change $\boldsymbol{\sigma} \to -\boldsymbol{\sigma}$. A second feature common to most (but not all) attractor neural networks is that *each* initial state will lead to pattern reconstruction, even nonsensical (random) ones.

The specific choice for the neural interactions, as made by Hopfield, is not optimal in the sense that, in addition to the desired stable states $\boldsymbol{\xi}^{\mu}$ and their mirror images $-\boldsymbol{\xi}^{\mu}$, even more unwanted spurious attractors are created. Yet this model will already push the analysis to the limits, as soon as we allow for the storage of an extensive number of patterns $\boldsymbol{\xi}^{\mu}$.

2.1 Hopfield Model Away from Saturation

The Hopfield model is obtained by generalising the recipe (2.1) to the case of having an arbitrary number p of patterns:

$$J_{ij}=\frac{1}{N}\sum_{\mu=1}^{p}\xi_i^\mu\xi_j^\mu, \qquad \theta_i=0 \tag{2.3}$$

with which the Hamiltonian H (1.26) (corresponding to sequential dynamics, upon eliminating self-interactions) and the pseudo-Hamiltonian $\tilde{H}$ (1.27) (corresponding to parallel dynamics) become

$$H(\boldsymbol{\sigma})=-\frac{1}{2}N\sum_{\mu=1}^{p}m_\mu^2(\boldsymbol{\sigma})+\frac{1}{2}p \tag{2.4}$$

$$\tilde{H}(\boldsymbol{\sigma})=-\frac{1}{\beta}\sum_{i=1}^{N}\log 2\cosh\left[\beta\sum_{\mu=1}^{p}\xi_i^\mu m_\mu(\boldsymbol{\sigma})\right] \tag{2.5}$$

with the so-called pattern overlaps

$$m_\mu(\boldsymbol{\sigma})=\frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i \tag{2.6}$$

Each of these p overlaps measures the resemblance between the current microscopic state $\boldsymbol{\sigma}$ and one particular pattern.

Equilibrium Order Parameter Equations. These can be obtained from the two free energies F and $\tilde{F}$:

$$F=-\frac{1}{\beta}\log\sum_{\boldsymbol{\sigma}}e^{-\beta H(\boldsymbol{\sigma})} \qquad \tilde{F}=-\frac{1}{\beta}\log\sum_{\boldsymbol{\sigma}}e^{-\beta\tilde{H}(\boldsymbol{\sigma})}$$

Upon introducing the short-hand notation $\boldsymbol{m}=(m_1,\ldots,m_p)$ and $\boldsymbol{\xi}_i=(\xi_i^1,\ldots,\xi_i^p)$, they can be expressed in terms of the density of states $\mathcal{D}(\boldsymbol{m})$:

$$\mathcal{D}(\boldsymbol{m})\equiv 2^{-N}\sum_{\boldsymbol{\sigma}}\delta\left[\boldsymbol{m}-\boldsymbol{m}(\boldsymbol{\sigma})\right]$$

$$F/N=-\frac{1}{\beta}\log 2-\frac{1}{\beta N}\log\int d\boldsymbol{m}\;\mathcal{D}(\boldsymbol{m})\;e^{\frac{1}{2}\beta N\boldsymbol{m}^2}+\frac{p}{2N} \tag{2.7}$$

$$\tilde{F}/N=-\frac{1}{\beta}\log 2-\frac{1}{\beta N}\log\int d\boldsymbol{m}\;\mathcal{D}(\boldsymbol{m})\;e^{\sum_{i=1}^{N}\log 2\cosh[\beta\boldsymbol{\xi}_i\cdot\boldsymbol{m}]} \tag{2.8}$$

In order to proceed we need to specify how the number of patterns p scales with the system size N. In this section we will assume p to be finite. One can now easily calculate the leading contribution to the density of states, using the integral representation of the δ-function and keeping in mind that according to (2.7,2.8) only terms exponential in N will retain statistical relevance for $N\to\infty$:

$$\lim_{N\to\infty}\frac{1}{N}\log\mathcal{D}(\boldsymbol{m})=\lim_{N\to\infty}\frac{1}{N}\log\int d\boldsymbol{x}\;e^{iN\boldsymbol{x}\cdot\boldsymbol{m}}\langle e^{-i\sum_{i=1}^{N}\sigma_i\boldsymbol{\xi}_i\cdot\boldsymbol{x}}\rangle_{\boldsymbol{\sigma}}$$

$$= \lim_{N\to\infty} \frac{1}{N} \log \int d\boldsymbol{x}\; e^{N\left[i\boldsymbol{x}\cdot\boldsymbol{m} + \langle \log\cos[\boldsymbol{\xi}\cdot\boldsymbol{x}]\rangle_{\boldsymbol{\xi}}\right]}$$

with the abbreviation

$$\langle \Phi(\boldsymbol{\xi})\rangle_{\boldsymbol{\xi}} = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \Phi(\boldsymbol{\xi}_i)$$

The leading contribution to both free energies can be expressed as a finite-dimensional integral, for large N dominated by the saddle-point that maximises the extensive exponent:

$$\lim_{N\to\infty} F/N = -\frac{1}{\beta N}\log\int d\boldsymbol{m}d\boldsymbol{x}\; e^{-N\beta f(\boldsymbol{m},\boldsymbol{x})} = \;\text{extr}\; f(\boldsymbol{m},\boldsymbol{x})$$

$$f(\boldsymbol{m},\boldsymbol{x}) = -\frac{1}{2}\boldsymbol{m}^2 - i\boldsymbol{x}\cdot\boldsymbol{m} - \frac{1}{\beta}\langle \log 2\cos\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{x}\right]\rangle_{\boldsymbol{\xi}}$$

and similarly

$$\lim_{N\to\infty} \tilde{F}/N = -\frac{1}{\beta N}\log\int d\boldsymbol{m}d\boldsymbol{x}\; e^{-N\beta \tilde{f}(\boldsymbol{m},\boldsymbol{x})} = \;\text{extr}\; \tilde{f}(\boldsymbol{m},\boldsymbol{x})$$

$$\tilde{f}(\boldsymbol{m},\boldsymbol{x}) = -\frac{1}{\beta}\langle \log 2\cosh\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}} - i\boldsymbol{x}\cdot\boldsymbol{m} - \frac{1}{\beta}\langle \log 2\cos\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{x}\right]\rangle_{\boldsymbol{\xi}}$$

The saddle-point equations for f and $\tilde{f}$ are given by:

$$\begin{array}{lll} f: & \boldsymbol{x} = i\boldsymbol{m} & i\boldsymbol{m} = \langle \boldsymbol{\xi}\tan\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{x}\right]\rangle_{\boldsymbol{\xi}} \\ \tilde{f}: & \boldsymbol{x} = i\langle \boldsymbol{\xi}\tanh\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}} & i\boldsymbol{m} = \langle \boldsymbol{\xi}\tan\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{x}\right]\rangle_{\boldsymbol{\xi}} \end{array}$$

In saddle points $\boldsymbol{x}$ turns out to be purely imaginary. After a shift of the integration contour we can eliminate $\boldsymbol{x}$ [1] and obtain

$$\begin{array}{ll} f: & \boldsymbol{m} = \langle \boldsymbol{\xi}\tanh\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}} \\ \tilde{f}: & \boldsymbol{m} = \langle \boldsymbol{\xi}\tanh\left[\beta\boldsymbol{\xi}\cdot\left[\langle \boldsymbol{\xi}'\tanh\left[\beta\boldsymbol{\xi}'\cdot\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}'}\right]\right]\rangle_{\boldsymbol{\xi}} \end{array}$$

The solutions of the above two equations will in general be identical. To see this, let us denote $\hat{\boldsymbol{m}} = \langle \boldsymbol{\xi}\tanh\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}}$, with which the saddle point equation for $\tilde{f}$ becomes:

$$\boldsymbol{m} = \langle \boldsymbol{\xi}\tanh\left[\beta\boldsymbol{\xi}\cdot\hat{\boldsymbol{m}}\right]\rangle_{\boldsymbol{\xi}} \qquad \hat{\boldsymbol{m}} = \langle \boldsymbol{\xi}\tanh\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}}$$

so

$$\left[\boldsymbol{m}-\hat{\boldsymbol{m}}\right]^2 = \langle \left[(\boldsymbol{\xi}\cdot\boldsymbol{m}) - (\boldsymbol{\xi}\cdot\hat{\boldsymbol{m}})\right]\left[\tanh(\beta\boldsymbol{\xi}\cdot\hat{\boldsymbol{m}}) - \tanh(\beta\boldsymbol{\xi}\cdot\boldsymbol{m})\right]\rangle_{\boldsymbol{\xi}}$$

The right-hand side can only be non-negative if $[\boldsymbol{m}-\hat{\boldsymbol{m}}]\cdot\boldsymbol{\xi} = 0$ for each $\boldsymbol{\xi}$ that contributes to the averages $\langle\ldots\rangle_{\boldsymbol{\xi}}$. For all choices of patterns where the covariance matrix $C_{\mu\nu} = \langle \xi_\mu\xi_\nu\rangle_{\boldsymbol{\xi}}$ is positive definite, we therefore obtain $\boldsymbol{m} = \hat{\boldsymbol{m}}$; as a consequence also the auxiliary saddle-point variables $\boldsymbol{x}$ will be identical. The occurrence of multiple saddle points corresponding

[1] In appendix ?? we discuss in more detail the technical details of how the choice of scaling with N of the auxiliary integration variables affects the saddle-point problem, and how to deal with complex saddle-points

to local minima of the free energy signals ergodicity breaking. Although among these only the *global* minimum will correspond to the thermodynamic equilibrium state, the non-global minima correspond to true ergodic components, i.e. on finite time-scales they will be just as relevant as the global minimum.

The final result is: for both types of dynamics (sequential and parallel) the overlap order parameters in the ergodic components are those solutions $\boldsymbol{m}^*$ of

$$\boldsymbol{m} = \langle \boldsymbol{\xi} \tanh [\beta \boldsymbol{\xi} \cdot \boldsymbol{m}] \rangle_{\boldsymbol{\xi}} \tag{2.9}$$

that (locally) minimise

$$f(\boldsymbol{m}) = \frac{1}{2}\boldsymbol{m}^2 - \frac{1}{\beta}\langle \log 2\cosh [\beta \boldsymbol{\xi} \cdot \boldsymbol{m}] \rangle_{\boldsymbol{\xi}} \tag{2.10}$$

The free energies of the ergodic compoments are

$$\lim_{N\to\infty} F/N = f(\boldsymbol{m}^*) \qquad\qquad \lim_{N\to\infty} \tilde{F}/N = 2f(\boldsymbol{m}^*) \tag{2.11}$$

Adding generating terms of the form $H \to H + \lambda g\left[\boldsymbol{m}(\boldsymbol{\sigma})\right]$ to the two Hamiltonians, for any function g, allows us identify

$$\langle g\left[\boldsymbol{m}(\boldsymbol{\sigma})\right]\rangle_{\rm eq} = \lim_{\lambda\to 0} \frac{\partial F}{\partial \lambda} = g\left[\boldsymbol{m}^*\right]$$

Consequently, in equilibrium the fluctuations in the overlap order parameters $\boldsymbol{m}(\boldsymbol{\sigma})$ (2.6) vanish in the limit $N \to \infty$, their deterministic values are simply given by $\boldsymbol{m}^*$.

Analysis of Order Parameter Equations: Mixture States. We will restrict our further discussion to the case of randomly drawn patterns, so

$$\langle \Phi(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} = 2^{-p} \sum_{\boldsymbol{\xi}\in\{-1,1\}^p} \Phi(\boldsymbol{\xi}), \qquad\qquad \langle \xi_\mu \xi_\nu \rangle_{\boldsymbol{\xi}} = \delta_{\mu\nu}$$

(generalisation to correlated patterns is in principle straightforward). We first establish an upper bound for the temperature $T = 1/\beta$ for non-trivial solutions $\boldsymbol{m}^*$ to exist, by writing (2.9) in integral form:

$$m_\mu = \beta \langle \xi_\mu \left(\boldsymbol{\xi} \cdot \boldsymbol{m}\right) \int_0^1 d\lambda \left[1 - \tanh^2\left(\beta\lambda \boldsymbol{\xi} \cdot \boldsymbol{m}\right)\right] \rangle_{\boldsymbol{\xi}}$$

from which we deduce

$$\begin{aligned} 0 &= \boldsymbol{m}^2 - \beta \langle (\boldsymbol{\xi} \cdot \boldsymbol{m})^2 \textstyle\int_0^1 d\lambda \left[1 - \tanh^2\left(\beta\lambda \boldsymbol{\xi} \cdot \boldsymbol{m}\right)\right] \rangle_{\boldsymbol{\xi}} \\ &\geq \boldsymbol{m}^2 - \beta \langle (\boldsymbol{\xi} \cdot \boldsymbol{m})^2 \rangle_{\boldsymbol{\xi}} = \boldsymbol{m}^2 \left[1 - \beta\right] \end{aligned}$$

For $T > 1$ the only solution of (2.9) is the paramagnetic state $\boldsymbol{m} = 0$ (which gives for the free energy per spin the standard results $-T \log 2$ and $-2T \log 2$, for sequential and parallel dynamics, respectively). At $T=1$, however, a continuous bifurcation occurs, which follows from expanding (2.9) for small $|\boldsymbol{m}|$ in powers of $\tau=\beta-1$:

$$\begin{aligned} m_\mu &= (1{+}\tau)m_\mu - \tfrac{1}{3}\textstyle\sum_{\nu\rho\lambda} m_\nu m_\rho m_\lambda \langle \xi_\mu\xi_\nu\xi_\rho\xi_\lambda\rangle_{\boldsymbol{\xi}} + \mathcal{O}(\boldsymbol{m}^5,\tau\boldsymbol{m}^3) \\ &= m_\mu\left[1+\tau-\boldsymbol{m}^2+\tfrac{2}{3}m_\mu^2\right]+\mathcal{O}(\boldsymbol{m}^5,\tau\boldsymbol{m}^3) \end{aligned}$$

The bifurcating saddle-point scales as $m_\mu=\tilde{m}_\mu\tau^{1/2}+\mathcal{O}(\tau^{3/2})$, with for each μ:

$$\tilde{m}_\mu=0 \quad \text{or} \quad 0=1-\tilde{\boldsymbol{m}}^2+\frac{2}{3}\tilde{m}_\mu^2$$

The solutions are of the form $\tilde{m}_\mu\in\{-\tilde{m},0,\tilde{m}\}$. If we denote with n the number of non-zero components in the vector $\tilde{\boldsymbol{m}}$, we derive from the above identities:

$$\tilde{m}_\mu=0 \quad \text{or} \quad \tilde{m}_\mu=\pm\left[\frac{3}{3n-2}\right]^{\frac{1}{2}}$$

These saddle point are called *mixture states*, since they correspond to microscopic configurations correlated equally with a finite number n of the stored patterns (or their negatives). Without loss of generality we can always perform gauge transformations on the set of stored patterns (permutations and reflections), such that these mixture states acquire the form

$$\boldsymbol{m}=m_n(\overbrace{1,\ldots,1}^{n \text{ times}},\overbrace{0,\ldots,0}^{p-n\text{ times}}) \qquad m_n=\left[\frac{3}{3n-2}\right]^{\frac{1}{2}}(\beta-1)^{1/2}+\ldots \tag{2.12}$$

These states are in fact saddle-points of the surface $f(\boldsymbol{m})$ (2.10) for any finite temperature, as can be verified by substitution of (2.12) as an *ansatz* into (2.9):

$$\begin{aligned} \mu\leq n: &\qquad m_n=\langle\xi_\mu\tanh\left[\beta m_n\textstyle\sum_{\nu\leq n}\xi_\nu\right]\rangle_{\boldsymbol{\xi}} \\ \mu>n: &\qquad 0=\langle\xi_\mu\tanh\left[\beta m_n\textstyle\sum_{\nu\leq n}\xi_\nu\right]\rangle_{\boldsymbol{\xi}} \end{aligned}$$

The second equation is automatically satisfied since the average factorises. The first equation leads to a condition determining the amplitude m_n of the mixture states:

$$m_n=\langle\left[\frac{1}{n}\sum_{\mu\leq n}\xi_\mu\right]\tanh\left[\beta m_n\sum_{\nu\leq n}\xi_\nu\right]\rangle_{\boldsymbol{\xi}} \tag{2.13}$$

The corresponding values of $f(\boldsymbol{m})$, to be denoted by f_n, are

$$f_n=\frac{1}{2}nm_n^2-\frac{1}{\beta}\langle\log 2\cosh\left[\beta m_n\sum_{\nu\leq n}\xi_\nu\right]\rangle_{\boldsymbol{\xi}} \tag{2.14}$$

The relevant question at this stage is whether these saddle points correspond to local minima of the surface $f(\boldsymbol{m})$ (2.10). The second derivative of $f(\boldsymbol{m})$ is given by

$$\frac{\partial^2 f(\boldsymbol{m})}{\partial m_\mu\partial m_\nu}=\delta_{\mu\nu}-\beta\langle\xi_\mu\xi_\nu\left[1-\tanh^2\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{m}\right]\right]\rangle_{\boldsymbol{\xi}} \tag{2.15}$$

(a local minimum corresponds to a positive definite second derivative). In the trivial saddle point $\boldsymbol{m} = 0$ this gives simply $\delta_{\mu\nu}[1-\beta]$, so at $T = 1$ this state destabilises. In a mixture state of the type (2.12) the second derivative becomes:

$$D^{(n)}_{\mu\nu} = \delta_{\mu\nu} - \beta\langle\xi_\mu\xi_\nu\left[1-\tanh^2\left[\beta m_n\sum_{\rho\leq n}\xi_\rho\right]\right]\rangle_{\boldsymbol{\xi}} \tag{2.16}$$

Due to the symmetries in the problem the spectrum of the matrix $D^{(n)}$ can be calculated. One finds three distinct eigenspaces:

	Eigenspace :	Eigenvalue :
I :	$\boldsymbol{x} = (0,\ldots,0,x_{n+1},\ldots,x_p)$	$1-\beta[1-Q]$
II :	$\boldsymbol{x} = (1,\ldots,1,0,\ldots,0)$	$1-\beta[1-Q+(1-n)R]$
III :	$\boldsymbol{x} = (x_1,\ldots,x_n,0,\ldots,0),\ \sum_\mu x_\mu = 0$	$1-\beta[1-Q+R]$

with

$$Q = \langle\tanh^2\left[\beta m_n \textstyle\sum_{\rho\leq n}\xi_\rho\right]\rangle_{\boldsymbol{\xi}}$$
$$R = \langle\xi_1\xi_2\tanh^2\left[\beta m_n \textstyle\sum_{\rho\leq n}\xi_\rho\right]\rangle_{\boldsymbol{\xi}}$$

Eigenspace *III* and the quantity R only come into play for $n > 1$. To find the smallest eigenvalue we need to know the sign of R. With the abbreviation $M_{\boldsymbol{\xi}} = \sum_{\rho\leq n}\xi_\rho$ we find:

$$\begin{aligned} n(n-1)R &= \langle M^2_{\boldsymbol{\xi}}\tanh^2\left[\beta m_n M_{\boldsymbol{\xi}}\right]\rangle_{\boldsymbol{\xi}} - n\langle\tanh^2\left[\beta m_n M_{\boldsymbol{\xi}}\right]\rangle_{\boldsymbol{\xi}} \\ &= \langle M^2_{\boldsymbol{\xi}}\tanh^2\left[\beta m_n M_{\boldsymbol{\xi}}\right]\rangle_{\boldsymbol{\xi}} - \langle M^2_{\boldsymbol{\xi}}\rangle_{\boldsymbol{\xi}}\langle\tanh^2\left[\beta m_n M_{\boldsymbol{\xi}}\right]\rangle_{\boldsymbol{\xi}} \\ &= \langle[M_{\boldsymbol{\xi}} - \langle M_{\boldsymbol{\xi}'}\rangle_{\boldsymbol{\xi}'}]^2\tanh^2\left[\beta m_n M_{\boldsymbol{\xi}}\right]\rangle_{\boldsymbol{\xi}} \geq 0 \end{aligned}$$

We may now identify the conditions for an n-mixture state to correspond to a local minimum of $f(\boldsymbol{m})$. For $n = 1$ the relevant eigenvalue is I, now the quantity Q simplifies considerably. For $n > 1$ the relevant eigenvalue is *III*, here we can combine Q and R into one single average (which reduces to a trivial expression for $n = 2$):

$$\begin{aligned} n = 1 :&\quad 1-\beta\left[1-\tanh^2[\beta m_1]\right] > 0 \\ n = 2 :&\quad 1-\beta > 0 \\ n \geq 3 :&\quad 1-\beta\left[1-\langle\tanh^2[\beta m_n \textstyle\sum_{\rho=3}^{n}\xi_\rho]\rangle_{\boldsymbol{\xi}}\right] > 0 \end{aligned}$$

The pure $n = 1$ states, correlated with one pattern only, are the desired solutions. They turn out to be stable for all $T < 1$, since partial differentiation with respect to β of the $n = 1$ amplitude equation (2.13) gives

$$m_1 = \tanh[\beta m_1] \quad\rightarrow\quad 1-\beta\left[1-\tanh^2[\beta m_1]\right] = \frac{m_1\left[1-\tanh^2[\beta m_1]\right]}{\partial m_1/\partial\beta}$$

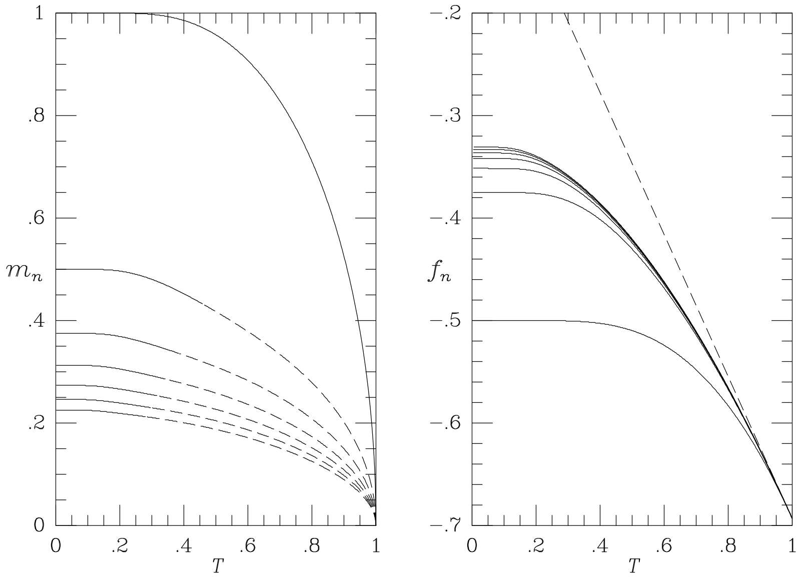


Figure 2.2: Left picture: Amplitudes m_n of the mixture states of the Hopfield model as a function of temperature. From top to bottom: $n = 1, 3, 5, 7, 9, 11, 13$. Solid: region where they are stable (i.e. local minima of f). Dashed: region where they are unstable. Right picture: corresponding 'free energies' f_n. From bottom to top: $n = 1, 3, 5, 7, 9, 11, 13$. Dashed line: 'free energy' of the paramagnetic state $m = 0$ (for comparison).

(clearly $\text{sgn}[m_1] = \text{sgn}[\partial m_1/\partial\beta]$). The $n = 2$ mixtures are always unstable. For $n \geq 3$ we have to solve the amplitude equations (2.13) numerically to evaluate their stability. The result is shown in figure 2.2, together with the corresponding 'free energies' f_n (2.14). It turns out that only for odd n will there be a critical temperature below which the n-mixture states are local minima of $f(\boldsymbol{m})$. From figure 2.2 we can also conclude that, in terms of the network functioning as an associative memory, noise is actually benificial in the sense that it can be used to eliminate the unwanted $n > 1$ ergodic components (whilst retaining the relevant ones: the pure $n = 1$ states).

In fact the overlap equations (2.9) do also allow for stable solutions different from the n-mixture states discussed here. They are in turn found to be continuously bifurcating mixtures of the mixture states. However, for random (or uncorrelated) patterns they come into existence only near $T = 0$ and play a marginal role; phase space is dominated by the odd n-mixture states.

Simulation Examples. We wil now illustrate with numerical simulations the functioning of the Hopfield model as an associative memory, and the description of the pattern recall process in terms of overlaps. Our system is an $N = 841$ Hopfield model, in which $p = 10$ patterns have been stored (see figure 2.3) according to the prescription (2.3). The two-dimensional arrangement of the spins in this example is just a guide to the eye, since the model is fully connected the spatial organisation of the spins in the network is physically irrelevant. The

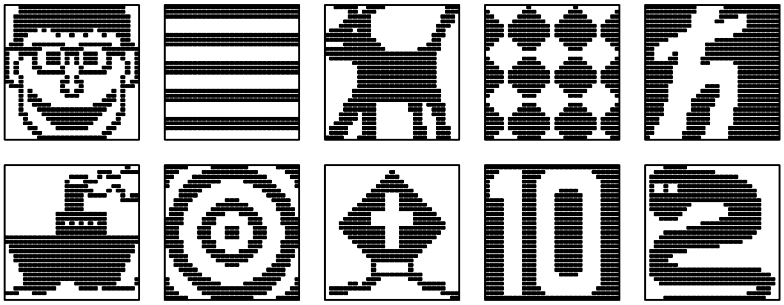


Figure 2.3: Information storage with the Hopfield model: $p = 10$ patterns represented as specific microscopic spin configurations in an $N = 841$ network.

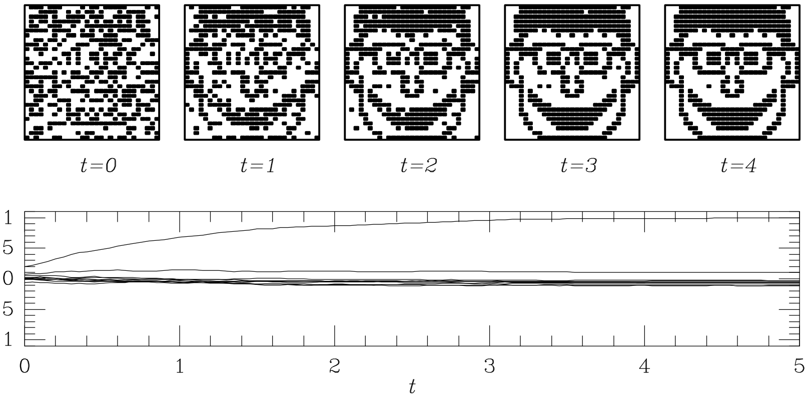


Figure 2.4: Dynamic reconstruction at $T = 0.1$ of a stored pattern from an initial state which is a corrupted version thereof. Top row: snapshots of the microscopic system state at times $t = 0, 1, 2, 3, 4$ iterations/spin. Bottom: the corresponding values of the $p = 10$ overlap order parameters as functions of time.

dynamics is sequential stochastic alignment to the local fields, as defined by the Glauber rule (1.5), for $T = 0.1$. In figure 2.4 we show the result of letting the system evolve in time from an initial state, which is a noisy version of one of the stored patterns (here 40% of the spins were flipped). The top row of graphs shows snapshots of the microscopic spin configuration as the system evolves stochastically in time. The bottom row shows the values of the $p = 10$ overlaps m_μ (defined in (2.6)), measured as functions of time; the one which evolves towards 1 correponding to the pattern being reconstructed. Figure 2.5 shows a similar experiment, here the initial state is simply drawn at random. The system subsequently evolves towards some mixture of the stored patterns. This mixture is not a symmetric one in the sense of the previous thermodynamic analysis, due to the fact that the patterns involved are significantly correlated (see figure 2.3).

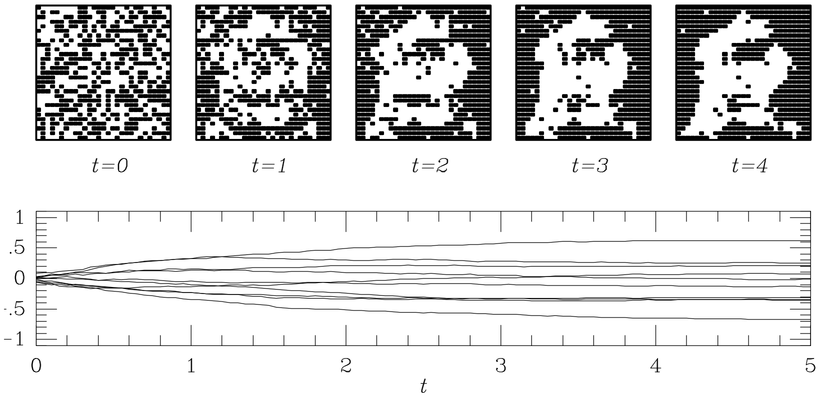


Figure 2.5: Evolution towards a spurious state at $T = 0.1$ from a randomly drawn initial state. Top row: snapshots of the microscopic system state at times $t = 0, 1, 2, 3, 4$ iterations/spin. Bottom: the corresponding values of the $p = 10$ overlap order parameters as functions of time.

2.2 Replica Analysis

The mean field analysis presented so far breaks down if p no longer remains finite for $N \to \infty$, but scales as $p = \alpha N$ $(\alpha > 0)$. The expressions (2.7,2.8) for the free energy per spin can no longer be evaluated by steepest descent, since the dimension of the integral involved diverges at the same time as the exponent of the integrand. The number of local minima of the Hamiltonians (2.4,2.5) and the number of ergodic components will diverge and we will encounter phenomena reminiscent of spin-glasses. As a consequence we will need corresponding methods of analysis, in the present case: replica theory. In fact for large $\alpha = p/N$ the interactions (2.3) of the Hopfield model will behave like independent Gaussian variables, which corresponds to the Sherrington-Kirkpatrick (SK) spin-glass model.

As an introduction to the problems associated with having an extremely large number of ergodic components and to the replica technique we will now discuss the equilibrium solution of the SK model. This model can alternatively be viewed as an Ising spin neural network with the (symmetric) interactions

$$J_{ij} = [1 - \delta_{ij}] \left[\frac{J_0}{N} + \frac{J}{\sqrt{N}} z_{ij} \right], \qquad \overline{z_{ij}} = 0,\ \overline{z_{ij}^2} = 1 \tag{2.17}$$

in which the z_{ij} $(i < j)$ are independent Gaussian random variables. We denote averaging over their distribution by $\overline{\cdots}$ (the factors in (2.17) involving N ensure appropriate scaling and statistical relevance of the two terms). In the context of spin-glasses the interactions represent quenched (frozen) disorder, we will forget about external fields θ_i for the moment. The Hamiltonian H (1.26), corresponding to sequential Glauber dynamics (1.5), becomes

$$H(\boldsymbol{\sigma}) = -\frac{1}{2} N J_0 m^2(\boldsymbol{\sigma}) + \frac{1}{2} J_0 - \frac{J}{\sqrt{N}} \sum_{i<j} \sigma_i \sigma_j z_{ij} \tag{2.18}$$

with the magnetization $m(\boldsymbol{\sigma}) = \frac{1}{N} \sum_k \sigma_k$. We clearly cannot calculate the free energy for every given realization of the interactions $\{J_{ij}\}$, furthermore it is to be expected that true macroscopic observables like the free energy and the magnitization only depend on the distribution of the interactions, not on their specific values.

Replica Calculation of the Disorder-Averaged Free Energy. We therefore average the free energy over the disorder distribution and concentrate on

$$\overline{F} = -\frac{1}{\beta}\overline{\log Z}, \qquad Z = \sum_{\boldsymbol{\sigma}} e^{-\beta H(\boldsymbol{\sigma})} \tag{2.19}$$

The average of the logarithm is transformed into an average of powers of the partition function Z, with the identity

$$\log Z = \lim_{n\to 0}\frac{1}{n}\left[Z^n - 1\right] \tag{2.20}$$

or, equivalently:

$$\overline{\log Z} = \lim_{n\to 0}\frac{1}{n}\log\overline{Z^n}$$

The so-called ‘replica trick' consists in evaluating the averages $\overline{Z^n}$ for integer values of n, and taking the limit $n \to 0$ afterwards under the assumption that the resulting expression is correct for non-integer values of n as well. The integer powers of Z are written as a product of terms, each of which can be interpreted as an equivalent copy, or 'replica' of the original system. The disorder-averaged free energy now becomes

$$\overline{F} = -\lim_{n\to 0}\frac{1}{\beta n}\left[\overline{Z^n} - 1\right] = -\lim_{n\to 0}\frac{1}{\beta n}\left[\sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n}\overline{e^{-\beta\sum_{\alpha=1}^{n}H(\boldsymbol{\sigma}^\alpha)}} - 1\right]$$

From now Roman indices will refer to sites, i.e. $i = 1\ldots N$, whereas Greek indices will refer to replicas, i.e. $\alpha = 1\ldots n$. We introduce the abbreviation for the Gaussian measure:

$$Dz = (2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}z^2}dz, \qquad \int Dz\; e^{xz} = e^{\frac{1}{2}x^2} \tag{2.21}$$

Upon insertion of the Hamiltonian (2.18) we obtain

$$\begin{aligned}\overline{F} &= -\tfrac{1}{\beta}N\log 2 - \lim_{n\to 0}\tfrac{1}{\beta n}\left[\langle e^{\frac{\beta J_0}{N}\sum_{i<j}\sum_\alpha \sigma_i^\alpha\sigma_j^\alpha}\prod_{i<j}\left[\int Dz\; e^{\frac{\beta J z}{\sqrt{N}}\sum_\alpha \sigma_i^\alpha\sigma_j^\alpha}\right]\rangle_{\{\boldsymbol{\sigma}^\alpha\}} - 1\right]\\ &= -\tfrac{1}{\beta}N\log 2 - \lim_{n\to 0}\tfrac{1}{\beta n}\left[\langle e^{\frac{\beta J_0}{2N}\sum_\alpha\sum_{i\neq j}\sigma_i^\alpha\sigma_j^\alpha + \frac{\beta^2J^2}{4N}\sum_{\alpha\gamma}\sum_{i\neq j}\sigma_i^\alpha\sigma_j^\alpha\sigma_i^\gamma\sigma_j^\gamma}\rangle_{\{\boldsymbol{\sigma}^\alpha\}} - 1\right]\end{aligned}$$

We now complete the sums over sites in this expression,

$$\sum_{i\neq j}\sigma_i^\alpha\sigma_j^\alpha = \left[\sum_i \sigma_i^\alpha\right]^2 - N, \qquad \sum_{i\neq j}\sigma_i^\alpha\sigma_j^\alpha\sigma_i^\gamma\sigma_j^\gamma = \left[\sum_i \sigma_i^\alpha\sigma_i^\gamma\right]^2 - N$$

The averaging over the spins $\{\boldsymbol{\sigma}^\alpha\}$ in our expression for $\overline{f}$ will now factorize nicely if we insert appropriate δ-functions (in their integral representations) to isolate the relevant terms, using

$$1 = \int d\boldsymbol{q}\;\prod_{\alpha\beta}\delta\left[q_{\alpha\beta} - \frac{1}{N}\sum_i \sigma_i^\alpha\sigma_i^\beta\right] = \left[\frac{N}{2\pi}\right]^{n^2}\int d\boldsymbol{q}d\hat{\boldsymbol{q}}\; e^{iN\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\left[q_{\alpha\beta} - \frac{1}{N}\sum_i \sigma_i^\alpha\sigma_i^\beta\right]}$$

$$1=\int d\boldsymbol{m}\;\prod_\alpha \delta\left[m_\alpha-\frac{1}{N}\sum_i \sigma_i^\alpha\right]=\left[\frac{N}{2\pi}\right]^n\int d\boldsymbol{m}d\hat{\boldsymbol{m}}\;e^{iN\sum_\alpha \hat{m}_\alpha\left[m_\alpha-\frac{1}{N}\sum_i\sigma_i^\alpha\right]}$$

The integrations are over the $n\times n$ matrices $\boldsymbol{q}$ and $\hat{\boldsymbol{q}}$ and over the n-vectors $\boldsymbol{m}$ and $\hat{\boldsymbol{m}}$. After inserting these integrals we obtain

$$\lim_{N\to\infty}\overline{F}/N=-\frac{1}{\beta}\log 2-\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{\beta N n}\left\{\left[\frac{N}{2\pi}\right]^{n^2+n}\int d\boldsymbol{q}d\hat{\boldsymbol{q}}d\boldsymbol{m}d\hat{\boldsymbol{m}}\right.$$

$$e^{N\left[i\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}q_{\alpha\gamma}+i\sum_\alpha \hat{m}_\alpha m_\alpha+\frac{1}{2}\beta J_0\sum_\alpha m_\alpha^2+\frac{1}{4}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}^2\right]}$$

$$\left.e^{-\frac{1}{2}n\beta J_0-\frac{1}{4}n^2\beta^2J^2}\langle e^{-i\sum_i\left[\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}\sigma_i^\alpha\sigma_i^\gamma+\sum_\alpha\hat{m}_\alpha\sigma_i^\alpha\right]}\rangle_{\{\boldsymbol{\sigma}^\alpha\}}-1\right\}$$

The spin averages factorise and are therefore reduced to single-site ones, involving only one n-replicated spin $(\sigma_1,\ldots,\sigma_n)$. Finally one assumes that the two limits $n\to 0$ and $N\to\infty$ commute. This allows us to evaluate the integral with the steepest-descent method, and subsequently evoke the initial identity (2.20) in 'reverse mode':

$$\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{Nn}\left[\int d\boldsymbol{x}\;e^{N\Phi(\boldsymbol{x})}-1\right]=\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{Nn}\left[e^{N\;\mathrm{extr}\Phi}-1\right]$$

$$=\lim_{N\to\infty}\lim_{n\to 0}\frac{1}{N}\log\;e^{N\;\mathrm{extr}\Phi/n}=\lim_{n\to 0}\frac{1}{n}\;\mathrm{extr}\Phi \tag{2.22}$$

The result of these manipulations is

$$\lim_{N\to\infty}\overline{F}/N=\lim_{n\to 0}\;\mathrm{extr}\;f(\boldsymbol{q},\boldsymbol{m};\hat{\boldsymbol{q}},\hat{\boldsymbol{m}}) \tag{2.23}$$

$$f(\boldsymbol{q},\boldsymbol{m};\hat{\boldsymbol{q}},\hat{\boldsymbol{m}})=-\frac{1}{\beta}\log 2-\frac{1}{\beta n}\left[\log\langle e^{-i\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}\sigma_\alpha\sigma_\gamma-i\sum_\alpha\hat{m}_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}\right.$$

$$\left.+i\sum_{\alpha\gamma}\hat{q}_{\alpha\gamma}q_{\alpha\gamma}+i\sum_\alpha\hat{m}_\alpha m_\alpha+\frac{1}{2}\beta J_0\sum_\alpha m_\alpha^2+\frac{1}{4}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}^2\right] \tag{2.24}$$

Variation of the parameters $\{q_{\alpha\beta}\}$ and $\{m_\alpha\}$ allows us to eliminate immediately the conjugate parameters $\{\hat{q}_{\alpha\beta}\}$ and $\{\hat{m}_\alpha\}$, since it leads to the saddle-point requirements

$$\hat{q}_{\alpha\beta}=\frac{1}{2}i\beta^2J^2q_{\alpha\beta}\qquad\qquad \hat{m}_\alpha=i\beta J_0m_\alpha \tag{2.25}$$

Upon elimination of $\{\hat{q}_{\alpha\beta},\hat{m}_\alpha\}$ according to (2.25) the result (2.23,2.24) is simplified to

$$\lim_{N\to\infty}\overline{F}/N=\lim_{n\to 0}\;\mathrm{extr}\;f(\boldsymbol{q},\boldsymbol{m}) \tag{2.26}$$

$$f(\boldsymbol{q},\boldsymbol{m})=-\frac{1}{\beta}\log 2+\frac{\beta J^2}{4n}\sum_{\alpha\gamma}q_{\alpha\gamma}^2+\frac{J_0}{2n}\sum_\alpha m_\alpha^2-\frac{1}{\beta n}\log\langle e^{\frac{1}{2}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma+\beta J_0\sum_\alpha m_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}} \tag{2.27}$$

Variation of the remaining parameters $\{q_{\alpha\beta}\}$ and $\{m_\alpha\}$ gives the final saddle-point equations

$$q_{\lambda\rho} = \frac{\langle\sigma_\lambda\sigma_\rho e^{\frac{1}{2}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma+\beta J_0\sum_\alpha m_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\frac{1}{2}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma+\beta J_0\sum_\alpha m_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}} \tag{2.28}$$

$$m_\lambda = \frac{\langle\sigma_\lambda e^{\frac{1}{2}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma+\beta J_0\sum_\alpha m_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\frac{1}{2}\beta^2J^2\sum_{\alpha\gamma}q_{\alpha\gamma}\sigma_\alpha\sigma_\gamma+\beta J_0\sum_\alpha m_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}} \tag{2.29}$$

The diagonal elements are always $q_{\alpha\alpha} = 1$. For high temperatures, $\beta = T^{-1} \to 0$, we obtain the trivial result

$$q_{\alpha\gamma} = \delta_{\alpha\gamma}, \qquad m_\alpha = 0$$

Assuming a continuous transition to a non-trivial state as the temperature is lowered, we can expand the saddle-point equations (2.28,2.29) in powers of $\boldsymbol{q}$ and $\boldsymbol{m}$ and look for bifurcations, which gives ($\lambda \neq \rho$):

$$q_{\lambda\rho} = \beta^2J^2q_{\lambda\rho} + \mathcal{O}(\boldsymbol{q},\boldsymbol{m})^2 \qquad\qquad m_\lambda = \beta J_0 m_\lambda + \mathcal{O}(\boldsymbol{q},\boldsymbol{m})^2$$

Therefore we expect second-order transitions either at $T = J_0$ (if $J_0 > J$) or at $T = J$ (if $J > J_0$). The remaining program is: find the saddle point $(\boldsymbol{q},\boldsymbol{m})$ for $T < \max\{J_0, J\}$ which for integer n minimises f, determine the corresponding minimum as a function of n, and finally take the limit $n \to 0$. This is in fact the most complicated part of the procedure.

Physical Interpretation of Saddle Points. To obtain a guide in how to select saddle-points we now turn to a different (although equivalent) version of the replica trick (2.20), which allows us to attach a physical meaning to the saddle-points $(\boldsymbol{m},\boldsymbol{q})$. This version transforms averages over a given measure W:

$$\begin{aligned}\frac{\sum_{\boldsymbol{\sigma}}\Phi(\boldsymbol{\sigma})W(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}}W(\boldsymbol{\sigma})} &= \frac{\sum_{\boldsymbol{\sigma}}\Phi(\boldsymbol{\sigma})W(\boldsymbol{\sigma})[\sum_{\boldsymbol{\sigma}}W(\boldsymbol{\sigma})]^{n-1}}{[\sum_{\boldsymbol{\sigma}}W(\boldsymbol{\sigma})]^n} \\ &= \lim_{n\to0}\sum_{\boldsymbol{\sigma}}\Phi(\boldsymbol{\sigma})W(\boldsymbol{\sigma})\,[\sum_{\boldsymbol{\sigma}}W(\boldsymbol{\sigma})]^{n-1} \\ &= \lim_{n\to0}\sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n}\Phi(\boldsymbol{\sigma}^1)\prod_{\alpha=1}^n W(\boldsymbol{\sigma}^\alpha)\end{aligned} \tag{2.30}$$

The trick again consists in evaluating this quantity for *integer* n, whereas the limit refers to non-integer n.

We use the above identity to write the distribution $P(m)$ of the magnetization in the SK model in equilibrium as

$$\begin{aligned}P(m) &= \frac{\sum_{\boldsymbol{\sigma}}\delta\left[m-\frac{1}{N}\sum_i\sigma_i\right]e^{-\beta H(\boldsymbol{\sigma})}}{\sum_{\boldsymbol{\sigma}}e^{-\beta H(\boldsymbol{\sigma})}} \\ &= \lim_{n\to0}\frac{1}{n}\sum_\gamma\sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n}\delta\left[m-\frac{1}{N}\sum_i\sigma_i^\gamma\right]\prod_\alpha e^{-\beta H(\boldsymbol{\sigma}^\alpha)}\end{aligned}$$

If we average this distribution over the disorder, we find identical expresions to those encountered in evaluating the disorder averaged free energy. By inserting the same delta-functions we arrive at the steepest descend integration (2.23) and find

$$\overline{P(m)} = \lim_{n\to 0} \frac{1}{n} \sum_{\gamma} \delta\left[m - m_\gamma\right] \tag{2.31}$$

where $\{m_\gamma\}$ refers to the relevant solution of (2.28,2.29).

Similarly we can imagine *two* systems $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ with identical realisation of the interactions $\{J_{ij}\}$, both in thermal equilibrium. We now use the replica identity to rewrite the distribution $P(q)$ for the mutual overlap between the microstates of the two systems

$$P(q) = \frac{\sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} \delta\left[q - \frac{1}{N}\sum_i \sigma_i \sigma_i'\right] e^{-\beta H(\boldsymbol{\sigma}) - \beta H(\boldsymbol{\sigma}')}}{\sum_{\boldsymbol{\sigma},\boldsymbol{\sigma}'} e^{-\beta H(\boldsymbol{\sigma}) - \beta H(\boldsymbol{\sigma}')}}$$

$$= \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{\lambda\neq\gamma} \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \delta\left[q - \frac{1}{N}\sum_i \sigma_i^\lambda \sigma_i^\gamma\right] \prod_\alpha e^{-\beta H(\boldsymbol{\sigma}^\alpha)}$$

Averaging over the disorder again leads to the steepest descend integration (2.23) and we find

$$\overline{P(q)} = \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{\lambda\neq\gamma} \delta\left[q - q_{\lambda\gamma}\right] \tag{2.32}$$

where $\{q_{\lambda\gamma}\}$ refers to the relevant solution of (2.28,2.29).

We can now partly interpret the saddlepoints $(\boldsymbol{m}, \boldsymbol{q})$, since the shape of $\overline{P(q)}$ and $\overline{P(m)}$ gives direct information on the structure of phase space with respect to ergodicity. The crucial observation is that for an ergodic mean-field system one always has

$$P(m) = \delta\left[m - \frac{1}{N}\sum_i \langle\sigma_i\rangle_{\rm eq}\right] \qquad P(q) = \delta\left[q - \frac{1}{N}\sum_i \langle\sigma_i\rangle_{\rm eq}^2\right] \tag{2.33}$$

If, on the other hand, there are L ergodic components in our system, each of which corresponding to a pure Gibbs state with microstate probabilities proportional to $\exp(-\beta H)$ and thermal averages $\langle\ldots\rangle_\ell$, and if we denote the probability of finding the system in component ℓ by W_ℓ, we find for $P(q)$ and $P(m)$:

$$P(m) = \sum_{\ell=1}^{L} W_\ell\ \delta\left[m - \frac{1}{N}\sum_i \langle\sigma_i\rangle_\ell\right] \qquad P(q) = \sum_{\ell,\ell'=1}^{L} W_\ell W_{\ell'}\ \delta\left[q - \frac{1}{N}\sum_i \langle\sigma_i\rangle_\ell \langle\sigma_i\rangle_{\ell'}\right]$$

For ergodic systems both $P(m)$ and $P(q)$ are δ-functions, for systems with a finite number of ergodic components they are composed of a finite sum of δ-functions. A diverging number of ergodic components, however, generally leads to distributions with continuous pieces. If we combine this interpretation with our results (2.31,2.31) we find that ergodicity is equivalent to the relevant saddle-point being of the form:

$$q_{\alpha\beta} = \delta_{\alpha\beta} + q\left[1 - \delta_{\alpha\beta}\right] \qquad m_\alpha = m \tag{2.34}$$

which is the so-called 'replica symmetry' (RS) ansatz. The physical meaning of m and q is provided by (2.33):

$$m=\frac{1}{N}\sum_i\overline{\langle\sigma_i\rangle_{\mathrm{eq}}}\qquad\qquad q=\frac{1}{N}\sum_i\overline{\langle\sigma_i\rangle_{\mathrm{eq}}^2}$$

Replica Symmetric Solution. Insertion of the above ansatz (2.34) into the equations (2.27,2.28,2.29) gives

$$f(\boldsymbol{q},\boldsymbol{m})=-\frac{1}{\beta}\log 2-\frac{1}{4}\beta J^2(1-q)^2+\frac{1}{2}J_0m^2-\frac{1}{\beta n}\log\langle e^{\frac{1}{2}q\beta^2J^2[\sum_\alpha\sigma_\alpha]^2+\beta J_0m\sum_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}+\mathcal{O}(n)$$

$$q=\frac{\langle\sigma_1\sigma_2e^{\frac{1}{2}q\beta^2J^2[\sum_\alpha\sigma_\alpha]^2+\beta J_0m\sum_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\frac{1}{2}q\beta^2J^2[\sum_\alpha\sigma_\alpha]^2+\beta J_0m\sum_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}\qquad\qquad m=\frac{\langle\sigma_1e^{\frac{1}{2}q\beta^2J^2[\sum_\alpha\sigma_\alpha]^2+\beta J_0m\sum_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\frac{1}{2}q\beta^2J^2[\sum_\alpha\sigma_\alpha]^2+\beta J_0m\sum_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}$$

We linearise the terms $[\sum_\alpha\sigma_\alpha]^2$ with the identity (2.21) and perform the average over the remaining spins. The solutions m and q turn out to be well defined for $n\to 0$ so we can take the limit:

$$\lim_{n\to 0}f(\boldsymbol{q},\boldsymbol{m})=-\frac{1}{\beta}\log 2-\frac{1}{4}\beta J^2(1-q)^2+\frac{1}{2}J_0m^2-\frac{1}{\beta}\int Dz\;\log\cosh\left[\beta J_0m+\beta Jz\sqrt{q}\right]\qquad(2.35)$$

$$q=\int Dz\;\tanh^2\left[\beta J_0m+\beta Jz\sqrt{q}\right]\qquad(2.36)$$

$$m=\int Dz\;\tanh\left[\beta J_0m+\beta Jz\sqrt{q}\right]\qquad(2.37)$$

Writing (2.37) in integral form gives

$$m=\beta J_0m\int_0^1 d\lambda\;\left[1-\int Dz\;\tanh^2\left[\lambda\beta J_0m+\beta Jz\sqrt{q}\right]\right]$$

From this expression, in combination with (2.36), we conclude:

$$\begin{array}{ll}T>J_0: & m=0\\ T>J_0\;\text{ and }\;T>J: & m=q=0\end{array}$$

Linearisation of (2.36,2.37) for small q and m shows the following continuous bifurcations:

	at	from	to
$J_0>J:$	$T=J_0$	$m=q=0$	$m\neq 0,\ q>0$
$J_0<J:$	$T=J$	$m=q=0$	$m=0,\ q>0$
$T<\max\{J_0,J\}:$	$T=J_0[1-q]$	$m=0,\ q>0$	$m\neq 0,\ q>0$

By solving numerically equations $T=J_0[1-q]$ and (2.37,2.36) we then arrive at the phase diagram shown in figure 2.6.

Breaking of Replica Symmetry: the AT Instability. If for the replica symmetric solution we calculate the entropy $S=\beta^2\partial F/\partial\beta$, we find that for small temperatures it becomes negative.

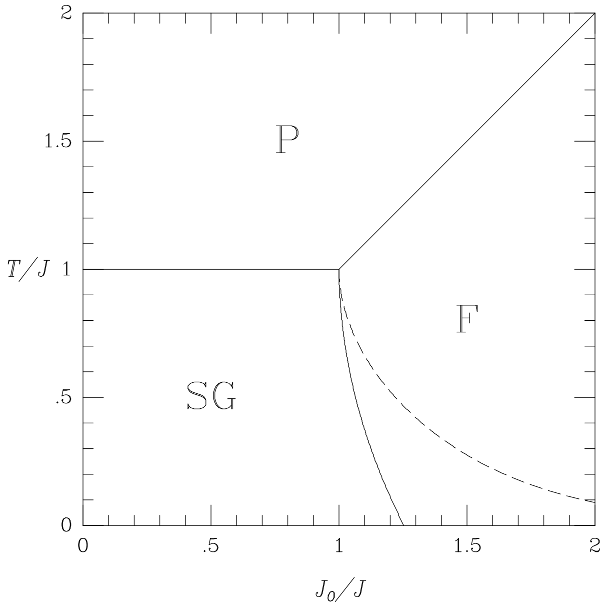


Figure 2.6: Phase diagram of the SK model, obtained from the replica-symmetric solution. P: paramagnetic phase, $m = q = 0$. SG: spin-glass phase, $m = 0$, $q \neq 0$. F: ferromagnetic phase, $m \neq 0$, $q \neq 0$. Solid lines: second-order transitions. Dashed: the AT instability.

One can easily prove that this cannot happen. First of all, straightforward differentiation shows

$$\frac{\partial S}{\partial \beta} = \beta \left[\langle H \rangle_{\rm eq}^2 - \langle H^2 \rangle_{\rm eq} \right] \leq 0$$

Let us now write $H(\boldsymbol{\sigma}) = H_0 + \hat{H}(\boldsymbol{\sigma})$, where H_0 is the ground state energy and $\hat{H}(\boldsymbol{\sigma}) \geq 0$ (zero only for ground state configurations, the number of which we denote by $N_0 \geq 1$). We now find

$$\lim_{\beta\to\infty} S = \lim_{\beta\to\infty} \left[\log \sum_{\boldsymbol{\sigma}} e^{-\beta \hat{H}(\boldsymbol{\sigma})} + \beta \langle \hat{H} \rangle_{\rm eq} \right] = \log N_0 \geq 0$$

We conclude that $S \geq 0$ for all β. At small temperatures the RS ansatz (2.34) is apparently incorrect in that it no longer corresponds to the minimum of $f(\boldsymbol{q}, \boldsymbol{m})$ (2.27).

If saddle-points without replica symmetry bifurcate continuously from the RS one, we can locate the occurrence of this 'replica symmetry breaking' (RSB) by studying the effect on $f(\boldsymbol{q}, \boldsymbol{m})$ of small fluctuations around the RS solution. It was shown by de Almeida and Thouless (1978) that the 'dangerous' fluctuations are of the form

$$q_{\alpha\beta} \to \delta_{\alpha\beta} + q\,[1 - \delta_{\alpha\beta}] + \eta_{\alpha\beta} \qquad \sum_{\beta} \eta_{\alpha\beta} = 0 \tag{2.38}$$

in which q is the solution of (2.36) and $\eta_{\alpha\beta} = \eta_{\beta\alpha}$. We now calculate the resulting change in $f(\boldsymbol{q}, \boldsymbol{m})$, away from the RS value $f(\boldsymbol{q}_{\rm RS}, \boldsymbol{m}_{\rm RS})$, the leading order of which is quadratic in the fluctuations $\{\eta_{\alpha\beta}\}$ since the RS solution (2.37,2.36) is a saddle-point:

$$f(\boldsymbol{q},\boldsymbol{m})-f(\boldsymbol{q}_{\rm RS},\boldsymbol{m}_{\rm RS})=\frac{\beta J^2}{4n}\sum_{\alpha\neq\gamma}\eta_{\alpha\gamma}^2-\frac{\beta^3J^4}{8n}\sum_{\alpha\neq\gamma}\sum_{\rho\neq\lambda}\eta_{\alpha\gamma}\eta_{\rho\lambda}G_{\alpha\gamma\rho\lambda}$$

with

$$G_{\alpha\gamma\rho\lambda}=\frac{\langle\sigma_\alpha\sigma_\gamma\sigma_\rho\sigma_\lambda e^{\frac{1}{2}q\beta^2J^2\left[\sum_\alpha\sigma_\alpha\right]^2+\beta mJ_0\sum_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\frac{1}{2}q\beta^2J^2\left[\sum_\alpha\sigma_\alpha\right]^2+\beta mJ_0\sum_\alpha\sigma_\alpha}\rangle_{\boldsymbol{\sigma}}}$$

Because of the index permutation symmetry in the spin-average we can write for $\alpha\neq\gamma$ and $\rho\neq\lambda$:

$$G_{\alpha\gamma\rho\lambda}=\delta_{\alpha\rho}\delta_{\gamma\lambda}+\delta_{\alpha\lambda}\delta_{\gamma\rho}+G_4\left[1-\delta_{\alpha\rho}\right]\left[1-\delta_{\gamma\lambda}\right]\left[1-\delta_{\alpha\lambda}\right]\left[1-\delta_{\gamma\rho}\right]$$

$$+\,G_2\left\{\delta_{\alpha\rho}\left[1-\delta_{\gamma\lambda}\right]+\delta_{\gamma\lambda}\left[1-\delta_{\alpha\rho}\right]+\delta_{\alpha\lambda}\left[1-\delta_{\gamma\rho}\right]+\delta_{\gamma\rho}\left[1-\delta_{\alpha\lambda}\right]\right\}$$

with

$$G_\ell=\frac{\int Dz\ \tanh^\ell\left[\beta J_0m+\beta Jz\sqrt{q}\right]\cosh^n\left[\beta J_0m+\beta Jz\sqrt{q}\right]}{\int Dz\ \cosh^n\left[\beta J_0m+\beta Jz\sqrt{q}\right]}$$

Only terms which involve precisely two δ-functions can contribute, because of the requirements $\alpha\neq\gamma$, $\rho\neq\lambda$ and $\sum_\beta\eta_{\alpha\beta}=0$. As a result:

$$f(\boldsymbol{q},\boldsymbol{m})-f(\boldsymbol{q}_{\rm RS},\boldsymbol{m}_{\rm RS})=\frac{\beta J^2}{4n}\left[1-\beta^2J^2\left(1-2G_2+G_4\right)\right]\sum_{\alpha\neq\gamma}\sum_{\rho\neq\lambda}\eta_{\alpha\gamma}^2$$

The condition for the RS solution to minimise $f(\boldsymbol{q},\boldsymbol{m})$, if compared to the so called 'replicon' fluctuations (2.38), is therefore

$$1>\beta^2J^2\lim_{n\to0}\left(1-2G_2+G_4\right)$$

After taking the limit in the expressions G_ℓ this condition can be written as

$$1>\beta^2J^2\int Dz\ \cosh^{-4}\left[\beta J_0m+\beta Jz\sqrt{q}\right]\tag{2.39}$$

The so-called AT line in the phase diagram where this condition ceases to be met, indicates a second-order transition to a spin-glass state where ergodicity is broken (i.e. the distribution $\overline{P(q)}$ (2.32) is no longer a δ-function). It is shown in figure 2.6 as a dashed line for $J_0/J>1$, and coincides with the line $T/J=1$ for $J_0<1$.

2.3 Sequential Hopfield Model Near Saturation

We now turn to the case where the number of patterns stored in the Hopfield model is extensive, i.e. $p=\alpha N$ in (2.3). Although we can still write the free energy in the form (2.7), this will not be of help since here it involves integrals over an extensive number of variables, so that steepest descent integration does not apply. Instead, following the approach of the

SK spin-glass model, we assume that we can average the free energy over the distribution of the patterns, with help of the replica-trick:

$$\overline{F} = -\lim_{n\to 0}\frac{1}{\beta n}\left[\sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n}\overline{e^{-\beta\sum_{\alpha=1}^{n}H(\boldsymbol{\sigma}^\alpha)}}-1\right]$$

Greek indices will denote either replica labels or pattern labels (it will be clear from the context), i.e. $\alpha,\beta=1,\ldots,n$ and $\mu,\nu=1,\ldots,p$. The $p\times N$ pattern components $\{\xi_i^\mu\}$ are assumed to be drawn independently at random from $\{-1,1\}$.

Replica Calculation of the Disorder-Averaged Free Energy. We first add to the Hamiltonian (2.4) a finite number ℓ of generating terms, that will allow us to obtain expectation values of the overlap order parameters m_μ (2.6) by differentiation of the free energy (since all patterns are equivalent in the calculation we may choose these ℓ nominated patterns arbitrarily):

$$H\to H+\sum_{\mu=1}^{\ell}\lambda_\mu\sum_i\sigma_i\xi_i^\mu \qquad \langle m_\mu(\boldsymbol{\sigma})\rangle_{\rm eq}=\lim_{\boldsymbol{\lambda}\to 0}\frac{\partial}{\partial\lambda_\mu}F/N \tag{2.40}$$

We know how to deal with a finite number of overlaps and corresponding patterns, therefore we average only over the disorder that is responsible for the complications: the patterns $\{\boldsymbol{\xi}^{\ell+1},\ldots,\boldsymbol{\xi}^p\}$ (as in the SK case we denote this disorder-averaging by $\overline{\cdots}$). Upon inserting the extended Hamiltonian into the replica-expression for the free energy, and assuming that the order of the limits $N\to\infty$ and $n\to 0$ can be interchanged, we obtain for large N:

$$\overline{F}/N=\frac{1}{2}\alpha-\frac{1}{\beta}\log 2-\lim_{n\to 0}\frac{1}{\beta Nn}\langle e^{-\beta\sum_{\mu\leq\ell}\sum_\alpha\left[\lambda_\mu\sum_i\sigma_i^\alpha\xi_i^\mu-\frac{1}{2N}[\sum_i\sigma_i^\alpha\xi_i^\mu]^2\right]}\overline{e^{\frac{\beta}{2N}\sum_\alpha\sum_{\mu>\ell}[\sum_i\sigma_i^\alpha\xi_i^\mu]^2}}\rangle_{\{\boldsymbol{\sigma}^\alpha\}}$$

We linearise the $\mu\leq\ell$ quadratic term using the identity (2.21), leading to $n\times\ell$ Gaussian integrals with $\boldsymbol{Dm}=(Dm_1^1,\ldots,Dm_n^\ell)$:

$$\overline{F}/N=\frac{1}{2}\alpha-\frac{1}{\beta}\log 2-\lim_{n\to 0}\frac{1}{\beta Nn}\int\boldsymbol{Dm}\langle e^{\sum_{\mu\leq\ell}\sum_\alpha\sum_i\sigma_i^\alpha\xi_i^\mu\left[\sqrt{\frac{\beta}{N}}m_\alpha^\mu-\beta\lambda_\mu\right]}\overline{e^{\frac{\beta}{2N}\sum_\alpha\sum_{\mu>\ell}[\sum_i\sigma_i^\alpha\xi_i^\mu]^2}}\rangle_{\{\boldsymbol{\sigma}^\alpha\}}$$

Anticipating that only terms exponential in the system size N will retain statistical relevance in the limit $N\to\infty$, we rescale the $n\times\ell$ integration variables $\boldsymbol{m}$ according to $\boldsymbol{m}\to\boldsymbol{m}\sqrt{\beta N}$:

$$\begin{aligned}\overline{F}/N=\frac{1}{2}\alpha-\frac{1}{\beta}\log 2-\lim_{n\to 0}\frac{1}{\beta Nn}\left[\frac{\beta N}{2\pi}\right]^{\frac{n\ell}{2}}\int d\boldsymbol{m}\;e^{-\frac{1}{2}\beta N\boldsymbol{m}^2}\times\\ \langle e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sum_i\sigma_i^\alpha\xi_i^\mu[m_\alpha^\mu-\lambda_\mu]}\;\overline{e^{\frac{\beta}{2N}\sum_\alpha\sum_{\mu>\ell}[\sum_i\sigma_i^\alpha\xi_i^\mu]^2}}\rangle_{\{\boldsymbol{\sigma}^\alpha\}}\end{aligned} \tag{2.41}$$

Next we turn to the disorder average, where we again linearise the exponent containing the pattern components using the identity (2.21), with $\boldsymbol{Dz}=(Dz_1,\ldots,Dz_n)$:

$$\overline{e^{\frac{\beta}{2N}\sum_\alpha\sum_{\mu>\ell}[\sum_i\sigma_i^\alpha\xi_i^\mu]^2}}=\left\{\overline{e^{\frac{1}{2}\sum_\alpha\left[(\frac{\beta}{N})^{\frac{1}{2}}\sum_i\sigma_i^\alpha\xi_i\right]^2}}\right\}^{p-\ell}=\left\{\int\boldsymbol{Dz}\;\overline{e^{(\frac{\beta}{N})^{\frac{1}{2}}\sum_\alpha z_\alpha\sum_i\sigma_i^\alpha\xi_i}}\right\}^{p-\ell}$$

$$= \left\{ \int D\boldsymbol{z} \, \prod_i \cosh\left[\left(\frac{\beta}{N}\right)^{\frac{1}{2}} \sum_\alpha z_\alpha \sigma_i^\alpha \right] \right\}^{p-\ell} = \left\{ \int D\boldsymbol{z} \; e^{\frac{\beta}{2N} \sum_{\alpha\beta} z_\alpha z_\beta \sum_i \sigma_i^\alpha \sigma_i^\beta + \mathcal{O}(\frac{1}{N})} \right\}^p \tag{2.42}$$

We are now as in the SK case led to introducing the replica order parameters $q_{\alpha\beta}$:

$$\begin{aligned} 1 &= \int d\boldsymbol{q} \;\; \prod_{\alpha\beta} \delta\left[q_{\alpha\beta} - \tfrac{1}{N} \sum_i \sigma_i^\alpha \sigma_i^\beta \right] \\ &= \left[\tfrac{N}{2\pi}\right]^{n^2} \int d\boldsymbol{q} d\hat{\boldsymbol{q}} \;\, e^{iN \sum_{\alpha\beta} \hat{q}_{\alpha\beta} \left[q_{\alpha\beta} - \frac{1}{N} \sum_i \sigma_i^\alpha \sigma_i^\beta \right]} \end{aligned}$$

Inserting (2.42) and the above identities into (2.41) and assuming that the limits $N \to \infty$ and $n \to 0$ commute gives:

$$\lim_{N\to\infty} \overline{F}/N = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \lim_{n\to 0} \frac{1}{\beta N n} \int d\boldsymbol{m} d\boldsymbol{q} d\hat{\boldsymbol{q}} \; e^{N\left[i \sum_{\alpha\beta} \hat{q}_{\alpha\beta} q_{\alpha\beta} - \frac{1}{2}\beta \boldsymbol{m}^2 + \alpha \log \int D\boldsymbol{z} \; e^{\frac{\beta}{2} \sum_{\alpha\beta} z_\alpha z_\beta q_{\alpha\beta}} \right]}$$

$$\times \langle e^{\beta \sum_{\mu\leq\ell} \sum_\alpha \sum_i \sigma_i^\alpha \xi_i^\mu [m_\alpha^\mu - \lambda_\mu] - i \sum_{\alpha\beta} \hat{q}_{\alpha\beta} \sum_i \sigma_i^\alpha \sigma_i^\beta} \rangle_{\{\boldsymbol{\sigma}^\alpha\}}$$

The n-dimensional Gaussian integral over $\boldsymbol{z}$ factorises in the standard way after appropriate rotation of the integration variables $\boldsymbol{z}$, with the result:

$$\log \int D\boldsymbol{z} \; e^{\frac{\beta}{2} \sum_{\alpha\beta} z_\alpha z_\beta q_{\alpha\beta}} = -\frac{1}{2} \log\det\left[\mathbf{I} - \beta \boldsymbol{q} \right]$$

in which $\mathbf{I}$ denotes the $n \times n$ identity matrix. The spin averages factorise and are reduced to single-site ones over the n-replicated spin $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$:

$$\lim_{N\to\infty} \overline{F}/N = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \lim_{n\to 0} \frac{1}{\beta N n} \int d\boldsymbol{m} d\boldsymbol{q} d\hat{\boldsymbol{q}} \; e^{N\left[i \sum_{\alpha\beta} \hat{q}_{\alpha\beta} q_{\alpha\beta} - \frac{1}{2}\beta \boldsymbol{m}^2 - \frac{1}{2}\alpha \log\det[\mathbf{I} - \beta \boldsymbol{q}] \right]}$$

$$\times \prod_i \langle e^{\beta \sum_{\mu\leq\ell} \sum_\alpha \sigma_\alpha \xi_i^\mu [m_\alpha^\mu - \lambda_\mu] - i \sum_{\alpha\beta} \hat{q}_{\alpha\beta} \sigma_\alpha \sigma_\beta} \rangle_{\boldsymbol{\sigma}}$$

and we arrive at integrals that can be evaluated by steepest descent, following the manipulations (2.22). If we denote averages over the remaining ℓ patterns in the familiar way

$$\boldsymbol{\xi} = (\xi_1, \ldots, \xi_\ell) \qquad \langle \Phi(\boldsymbol{\xi}) \rangle_{\boldsymbol{\xi}} = 2^{-\ell} \sum_{\boldsymbol{\xi} \in \{-1,1\}^\ell} \Phi(\boldsymbol{\xi})$$

we can write the final result in the form

$$\lim_{N\to\infty} \overline{F}/N = \lim_{n\to 0} \;\text{extr}\; f(\boldsymbol{m}, \boldsymbol{q}, \hat{\boldsymbol{q}}) \tag{2.43}$$

$$f(\boldsymbol{m}, \boldsymbol{q}, \hat{\boldsymbol{q}}) = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \frac{1}{\beta n} \left[\langle \log \langle e^{\beta \sum_{\mu\leq\ell} \sum_\alpha \sigma_\alpha \xi_\mu [m_\alpha^\mu - \lambda_\mu] - i \sum_{\alpha\beta} \hat{q}_{\alpha\beta} \sigma_\alpha \sigma_\beta} \rangle_{\boldsymbol{\sigma}} \rangle_{\boldsymbol{\xi}} \right.$$

$$\left. + i \sum_{\alpha\beta} \hat{q}_{\alpha\beta} q_{\alpha\beta} - \frac{1}{2}\beta \boldsymbol{m}^2 - \frac{1}{2}\alpha \log\det\left[\mathbf{I} - \beta \boldsymbol{q} \right] \right]$$

Having arrived at a saddle-point problem we now first identify the expectation values of the overlaps with (2.40) (note: extremisation with respect to the saddle-point variables and differentiation with respect to $\boldsymbol{\lambda}$ commute):

$$\overline{\langle m_\mu(\boldsymbol{\sigma})\rangle_{\rm eq}} = \lim_{n\to 0}\lim_{\boldsymbol{\lambda}\to 0}\frac{\partial}{\partial\lambda_\mu}\ \text{extr}\ f(\boldsymbol{m},\boldsymbol{q},\hat{\boldsymbol{q}})$$

$$= \lim_{n\to 0}\langle\xi_\mu\frac{\langle\frac{1}{n}\sum_\alpha\sigma_\alpha e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}\rangle_{\boldsymbol{\sigma}}}\rangle_{\boldsymbol{\xi}} \tag{2.44}$$

which is to be evaluated in the $\boldsymbol{\lambda}=0$ saddle-point. Having served their purpose, the generating fields λ_μ can be set to zero and we can restrict ourselves to the $\boldsymbol{\lambda}=0$ saddle-point problem:

$$f(\boldsymbol{m},\boldsymbol{q},\hat{\boldsymbol{q}}) = \frac{1}{2}\alpha - \frac{1}{\beta}\log 2 - \frac{1}{\beta n}\left[i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}q_{\alpha\beta} - \frac{1}{2}\beta\boldsymbol{m}^2 - \frac{1}{2}\alpha\log\det\left[\mathbb{I}-\beta\boldsymbol{q}\right]\right.$$

$$\left. +\langle\log\langle e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}\rangle_{\boldsymbol{\sigma}}\rangle_{\boldsymbol{\xi}}\right] \tag{2.45}$$

Variation of the parameters $\{m_\alpha^\mu,\ q_{\alpha\beta},\ \hat{q}_{\alpha\beta}\}$ gives the saddle-point equations:

$$m_\alpha^\mu = \langle\xi_\mu\frac{\langle\sigma_\alpha e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}\rangle_{\boldsymbol{\sigma}}}\rangle_{\boldsymbol{\xi}} \tag{2.46}$$

$$q_{\lambda\rho} = \langle\frac{\langle\sigma_\lambda\sigma_\rho e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\sum_{\mu\leq\ell}\sum_\alpha\sigma_\alpha\xi_\mu m_\alpha^\mu - i\sum_{\alpha\beta}\hat{q}_{\alpha\beta}\sigma_\alpha\sigma_\beta}\rangle_{\boldsymbol{\sigma}}}\rangle_{\boldsymbol{\xi}} \tag{2.47}$$

$$\hat{q}_{\lambda\rho} = \frac{1}{2}i\alpha\beta\frac{\int d\boldsymbol{z}\ z_\lambda z_\rho e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbb{I}-\beta\boldsymbol{q}]\boldsymbol{z}}}{\int d\boldsymbol{z}\ e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbb{I}-\beta\boldsymbol{q}]\boldsymbol{z}}} \tag{2.48}$$

furthermore,

$$\overline{\langle m_\mu(\boldsymbol{\sigma})\rangle_{\rm eq}} = \lim_{n\to 0}\frac{1}{n}\sum_\alpha m_\alpha^\mu \tag{2.49}$$

replaces the identification (2.44). As expected, one always has $q_{\alpha\alpha}=1$. The diagonal elements $\hat{q}_{\alpha\alpha}$ drop out of (2.46,2.47), their values are simply given as functions of the remaining parameters by (2.48).

Physical Interpretation of Saddle Points. We proceed along the lines of the SK model. If we apply the alternative version (2.30) of the replica trick to the Hopfield model, we can write the distribution of the ℓ overlaps $\boldsymbol{m}=(m_1,\ldots,m_\ell)$ in equilibrium as

$$P(\boldsymbol{m}) = \lim_{n\to 0}\frac{1}{n}\sum_\gamma\sum_{\boldsymbol{\sigma}^1\ldots\boldsymbol{\sigma}^n}\delta\left[\boldsymbol{m}-\frac{1}{N}\sum_i\sigma_i^\gamma\boldsymbol{\xi}_i\right]\prod_\alpha e^{-\beta H(\boldsymbol{\sigma}^\alpha)}$$

with $\boldsymbol{\xi}_i = (\xi_i^1, \ldots, \xi_i^\ell)$. Averaging this distribution over the disorder leads to expressions identical to those encountered in evaluating the disorder averaged free energy. By inserting the same delta-functions we arrive at the steepest descend integration (2.43,2.45) and find

$$\overline{P(\boldsymbol{m})} = \lim_{n\to 0} \frac{1}{n} \sum_{\gamma} \delta\left[\boldsymbol{m} - \boldsymbol{m}_\gamma\right] \tag{2.50}$$

where $\boldsymbol{m}_\gamma = (m_\gamma^1, \ldots, m_\gamma^\ell)$ refers to the relevant solution of (2.46,2.47,2.48).

Similarly we imagine two systems $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ with identical realisation of the interactions $\{J_{ij}\}$, both in thermal equilibrium, and use (2.30) to rewrite the distribution $P(q)$ for the mutual overlap between the microstates of the two systems

$$P(q) = \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{\lambda\neq\gamma} \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \delta\left[q - \frac{1}{N}\sum_i \sigma_i^\lambda \sigma_i^\gamma\right] \prod_\alpha e^{-\beta H(\boldsymbol{\sigma}^\alpha)}$$

Averaging over the disorder again leads to the steepest descend integration (2.43,2.45) and we find

$$\overline{P(q)} = \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{\lambda\neq\gamma} \delta\left[q - q_{\lambda\gamma}\right] \tag{2.51}$$

where $\{q_{\lambda\gamma}\}$ refers to the relevant solution of (2.46,2.47,2.48).

Finally we analyse the physical meaning of the conjugate parameters $\{\hat{q}_{\alpha\beta}\}$ for $\alpha \neq \beta$. We will do this in more detail, it being rather specific for the Hopfield model and slightly different from the derivations above. Again we imagine two systems $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ with identical interactions $\{J_{ij}\}$, both in thermal equilibrium. We now use (2.30) to evaluate the covariance of the overlaps corresponding to non-nominated patterns:

$$r = \frac{1}{\alpha} \sum_{\mu=\ell+1}^{p} \overline{\langle \frac{1}{N}\sum_i \sigma_i \xi_i^\mu \rangle_{\rm eq} \langle \frac{1}{N}\sum_i \sigma_i' \xi_i^\mu \rangle_{\rm eq}} \tag{2.52}$$

$$= \lim_{n\to 0} \frac{N-\ell/\alpha}{n(n-1)} \sum_{\lambda\neq\gamma} \sum_{\boldsymbol{\sigma}^1 \ldots \boldsymbol{\sigma}^n} \overline{\left[\frac{1}{N}\sum_i \sigma_i^\lambda \xi_i^p\right] \left[\frac{1}{N}\sum_i \sigma_i^\gamma \xi_i^p\right] \prod_\alpha e^{-\beta H(\boldsymbol{\sigma}^\alpha)}}$$

(using the equivalence of all such patterns). We next perform the same manipulations as in calculating the free energy. Here the disorder average involves

$$\overline{\left[\frac{1}{\sqrt{N}}\sum_i \sigma_i^\lambda \xi_i^p\right] \left[\frac{1}{\sqrt{N}}\sum_i \sigma_i^\gamma \xi_i^p\right] e^{\frac{\beta}{2N}\sum_\alpha \sum_{\mu>\ell} \left[\sum_i \sigma_i^\alpha \xi_i^\mu\right]^2}}$$

$$= \left\{ \int D\boldsymbol{z}\ \overline{e^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}} \sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}} \right\}^{p-\ell-1} \int \frac{D\boldsymbol{z}}{\beta} \frac{\partial^2}{\partial z_\lambda \partial z_\gamma} \overline{e^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}} \sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}}$$

$$= \left\{ \int D\boldsymbol{z}\ \overline{e^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}} \sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}} \right\}^{p-\ell-1} \int D\boldsymbol{z} \frac{z_\lambda z_\gamma}{\beta} \overline{e^{\left(\frac{\beta}{N}\right)^{\frac{1}{2}} \sum_\alpha z_\alpha \sum_i \sigma_i^\alpha \xi_i}}$$

(after partial integration). We finally obtain an expression which involves the surface (2.45):

$$r = \frac{1}{\beta} \lim_{n\to 0} \frac{1}{n(n-1)} \sum_{\lambda\neq\rho} \lim_{N\to\infty} \frac{\int d\boldsymbol{m} d\boldsymbol{q} d\hat{\boldsymbol{q}} \left[\frac{\int d\boldsymbol{z}\ z_\lambda z_\rho\ e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbb{1}-\beta\boldsymbol{q}]\boldsymbol{z}}}{\int d\boldsymbol{z}\ e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbb{1}-\beta\boldsymbol{q}]\boldsymbol{z}}}\right] e^{-\beta n N f(\boldsymbol{m},\boldsymbol{q},\hat{\boldsymbol{q}})}}{\int d\boldsymbol{m} d\boldsymbol{q} d\hat{\boldsymbol{q}}\ e^{-\beta n N f(\boldsymbol{m},\boldsymbol{q},\hat{\boldsymbol{q}})}}$$

The normalisation of the above integral over $\{\boldsymbol{m},\boldsymbol{q},\hat{\boldsymbol{q}}\}$ follows from using the replica procedure to rewrite unity. The integration being dominated by the minima of f, we can use the saddle-point equations (2.48) to arrive at

$$\lim_{n\to 0}\frac{1}{n(n-1)}\sum_{\lambda\neq\rho}\hat{q}_{\lambda\rho}=\frac{1}{2}i\alpha\beta^2 r \tag{2.53}$$

The result (2.52,2.53) provides a physical interpretation of the order parameters $\{\hat{q}_{\alpha\beta}\}$.

Ergodicity implies that the distributions $\overline{P(q)}$ and $\overline{P(\boldsymbol{m})}$ are δ-functions, this is equivalent to the relevant saddle-point being of the form:

$$m^\mu_\alpha=m_\mu \qquad q_{\alpha\beta}=\delta_{\alpha\beta}+q\,[1-\delta_{\alpha\beta}] \qquad \hat{q}_{\alpha\beta}=\frac{1}{2}i\alpha\beta^2\,[R\delta_{\alpha\beta}+r\,[1-\delta_{\alpha\beta}]] \tag{2.54}$$

which is the 'replica symmetry' (RS) ansatz for the Hopfield model. The RS form for $\{q_{\alpha\beta}\}$ and $\{m^\mu_\alpha\}$ is a direct consequence of the corresponding distributions being δ-functions, whereas the RS form for $\{\hat{q}_{\alpha\beta}\}$ subsequently follows from (2.48). The physical meaning of m_μ and q is

$$m_\mu=\overline{\langle m_\mu(\boldsymbol{\sigma})\rangle_{\rm eq}} \qquad q=\frac{1}{N}\sum_i\overline{\langle\sigma_i\rangle^2_{\rm eq}}$$

Before proceeding with a full analysis of the RS saddle-point equations, we finally make a few tentative statements on the phase diagram. For $\beta=0$ we obtain the trivial result $q_{\lambda\rho}=\delta_{\lambda\rho}$, $\hat{q}_{\lambda\rho}=0$, $m^\mu_\alpha=0$. We can identify continuous bifurcations to a non-trivial state by expanding the saddle-point equations in first order in the relevant parameters:

$$m^\mu_\alpha=\beta m^\mu_\alpha+\ldots \qquad q_{\lambda\rho}=-2i\hat{q}_{\lambda\rho}+\ldots \quad (\lambda\neq\rho)$$

$$\hat{q}_{\lambda\rho}=\frac{1}{2}i\frac{\alpha\beta}{1-\beta}\left[\delta_{\lambda\rho}+\frac{\beta}{1-\beta}q_{\lambda\rho}\,[1-\delta_{\lambda\rho}]\right]+\ldots$$

By combining the equations for $\boldsymbol{q}$ and $\hat{\boldsymbol{q}}$ we obtain

$$q_{\lambda\rho}=\alpha\left[\frac{\beta}{1-\beta}\right]^2 q_{\lambda\rho}+\ldots$$

Therefore we expect a second order transition at $T=1+\sqrt{\alpha}$ from the trivial state to an ordered state where $\langle m_\mu\rangle_{\rm eq}=0$ (a spin-glass state).

2.4 Replica Symmetric Solution

The symmetry of the RS ansatz (2.54) for the saddle-point allows us to diagonalise the maxtrix $\boldsymbol{\Lambda}=\mathbf{I}-\beta\boldsymbol{q}$ which we encountered in the saddle-point problem:

$$\Lambda_{\alpha\beta}=[1-\beta(1-q)]\,\delta_{\alpha\beta}-\beta q$$

Eigenspace :	Eigenvalue :	Multiplicity :
$\boldsymbol{x}=(1,\ldots,1)$	$1-\beta(1-q)-\beta qn$	1
$\sum_\alpha x_\alpha=0$	$1-\beta(1-q)$	$n-1$

so that

$$\begin{aligned}\log\det\boldsymbol{\Lambda} &= \log\left[1-\beta(1-q)-\beta qn\right]+(n-1)\log\left[1-\beta(1-q)\right]\\ &= n\left[\log\left[1-\beta(1-q)\right]-\frac{\beta q}{1-\beta(1-q)}\right]+\mathcal{O}(n^2)\end{aligned}$$

Inserting the RS ansatz (2.54) for the saddle-point into (2.45), utilising the above expression for the determinant and the short-hand $\boldsymbol{m}=(m_1,\ldots,m_\ell)$, gives

$$\begin{aligned}f(\boldsymbol{m}_{\mathrm{RS}},\boldsymbol{q}_{\mathrm{RS}},\hat{\boldsymbol{q}}_{\mathrm{RS}}) &= -\frac{1}{\beta}\log 2+\frac{1}{2}\alpha\left[1+\beta r(1-q)\right]+\frac{1}{2}\boldsymbol{m}^2+\frac{\alpha}{2\beta}\left[\log\left[1-\beta(1-q)\right]-\frac{\beta q}{1-\beta(1-q)}\right]\\ &\quad -\frac{1}{\beta n}\langle\log\langle e^{\beta\boldsymbol{m}\cdot\boldsymbol{\xi}\sum_\alpha\sigma_\alpha+\frac{1}{2}\alpha r\beta^2\left[\sum_\alpha\sigma_\alpha\right]^2}\rangle_{\boldsymbol{\sigma}}\rangle_{\boldsymbol{\xi}}+\mathcal{O}(n)\end{aligned}$$

We now linearise the squares in the spin averages with (2.21), subsequently average over the replicated spin $\boldsymbol{\sigma}$, use $\cosh^n[x]=1+n\log\cosh[x]+\mathcal{O}(n^2)$ and take the limit $n\to 0$:

$$\begin{aligned}\lim_{N\to\infty}\overline{F}_{\mathrm{RS}}/N &= \lim_{n\to 0}f(\boldsymbol{m}_{\mathrm{RS}},\boldsymbol{q}_{\mathrm{RS}},\hat{\boldsymbol{q}}_{\mathrm{RS}})\\ &= \frac{1}{2}\boldsymbol{m}^2+\frac{1}{2}\alpha\left[1+\beta r(1-q)+\frac{1}{\beta}\log\left[1-\beta(1-q)\right]-\frac{q}{1-\beta(1-q)}\right]\\ &\quad -\frac{1}{\beta}\langle\int Dz\ \log 2\cosh\beta\left[\boldsymbol{m}\cdot\boldsymbol{\xi}+z\sqrt{\alpha r}\right]\rangle_{\boldsymbol{\xi}}\end{aligned}\tag{2.55}$$

The saddle point equations for $\boldsymbol{m}$, q and r can be obtained either by insertion of the RS ansatz (2.54) into (2.46,2.47,2.48) and subsequently taking the $n\to 0$ limit, or by variation of the RS expression (2.55). The latter route is the fastest one. After performing partial integrations where appropriate we obtain the final result:

$$\boldsymbol{m}=\langle\boldsymbol{\xi}\int Dz\ \tanh\beta\left[\boldsymbol{m}\cdot\boldsymbol{\xi}+z\sqrt{\alpha r}\right]\rangle_{\boldsymbol{\xi}}\tag{2.56}$$

$$q=\langle\int Dz\ \tanh^2\beta\left[\boldsymbol{m}\cdot\boldsymbol{\xi}+z\sqrt{\alpha r}\right]\rangle_{\boldsymbol{\xi}}\tag{2.57}$$

$$r=\frac{q}{\left[1-\beta(1-q)\right]^2}\tag{2.58}$$

By substitution of (2.58) into the remaining equations this set can easily be further reduced, should the need arise. In case of multiple solutions of (2.56,2.58,2.57) the relevant saddle point is the one that minimises (2.55). Clearly for $\alpha=0$ we recover our previous results (2.9,2.10).

Analysis of RS Order Parameter Equations and Phase Diagram. We first establish an upper bound for the temperature $T=1/\beta$ for non-trivial solutions of the set (2.56,2.57,2.58) to exist, by writing (2.56) in integral form:

$$m_\mu=\beta\langle\xi_\mu\left(\boldsymbol{\xi}\cdot\boldsymbol{m}\right)\int_0^1 d\lambda\int Dz\left[1-\tanh^2\beta\left(\lambda\boldsymbol{\xi}\cdot\boldsymbol{m}+z\sqrt{\alpha r}\right)\right]\rangle_{\boldsymbol{\xi}}$$

from which we deduce

$$\begin{aligned} 0 &= \boldsymbol{m}^2 - \beta\langle(\boldsymbol{\xi}\cdot\boldsymbol{m})^2 \int_0^1 d\lambda \int Dz \left[1 - \tanh^2\beta\,(\lambda\boldsymbol{\xi}\cdot\boldsymbol{m} + z\sqrt{\alpha r})\right]\rangle_{\boldsymbol{\xi}} \\ &\geq \boldsymbol{m}^2 - \beta\langle(\boldsymbol{\xi}\cdot\boldsymbol{m})^2\rangle_{\boldsymbol{\xi}} = \boldsymbol{m}^2\,[1-\beta] \end{aligned}$$

Therefore $\boldsymbol{m} = 0$ for $T > 1$. If $T > 1$ we obtain in turn from (2.57,2.58), using $\tanh^2(x) \leq x^2$ and $0 \leq q \leq 1$:

$$q = 0 \qquad \text{or} \qquad q \leq 1 + \sqrt{\alpha} - T$$

We conclude that $q = 0$ for $T > 1 + \sqrt{\alpha}$. Secondly, for the free energy (2.10) to be well defined we must require $q > 1 - T$. Linearisation of (2.56,2.57) for small q and $\boldsymbol{m}$ shows the continuous bifurcations:

	at	from	to
$\alpha > 0:$	$T = 1 + \sqrt{\alpha}$	$\boldsymbol{m} = 0,\ q = 0$	$\boldsymbol{m} = 0,\ q > 0$
$\alpha = 0:$	$T = 1$	$\boldsymbol{m} = 0,\ q = 0$	$\boldsymbol{m} \neq 0,\ q > 0$

The upper bound $T = 1 + \sqrt{\alpha}$ turns out to be the critical temperature indicating (for $\alpha > 0$) a second order transition to a spin-glass state, where there is no significant alignment of the spins in the direction of one particular pattern, but still a certain degree of local freezing. Since $\boldsymbol{m} = 0$ for $T > 1$ this spin-glass state persists at least down to $T = 1$. The quantitative details of the spin-glass state are obtained by inserting $\boldsymbol{m} = 0$ into (2.57,2.58) (since (2.56) is fulfilled automatically).

The impact on the saddle-point equations (2.56,2.57) of having $\alpha > 0$, a smoothening of the hyperbolic tangent by convolution with a Gaussian kernel, can be viewed as noise caused by interference between the attractors. The natural strategy for solving (2.56,2.57) is therefore to make an ansatz for the nominated overlaps $\boldsymbol{m}$ of the type (2.12) (the mixture states). Insertion of this ansatz into the saddle-point equations indeed leads to self-consistent solutions. One can solve numerically the remaining equations for the amplitudes of the mixture states and evaluate their stability by calculating the eigenvalues of the second derivative of $f(\boldsymbol{m}, \boldsymbol{q}, \hat{\boldsymbol{q}})$, in the same way as for $\alpha = 0$. The calculations are just more involved. It then turns out that even mixtures are again unstable for any T and α, whereas odd mixtures can become locally stable for sufficiently small T and α. Among the mixture states, the pure states, where the vector $\boldsymbol{m}$ has only one nonzero component, are the first to stabilise as the temperature is lowered. These pure states, together with the spin-glass state ($\boldsymbol{m} = 0,\ q > 0$), we will study in more detail.

Let us first calculate the second derivatives of (2.55) and evaluate them in the spin-glass saddle-point. One finds, after elimination of r with (2.58):

$$\partial^2 f/\partial m_\mu \partial m_\nu = \delta_{\mu\nu}\,[1-\beta(1-q)] \qquad\qquad \partial^2 f/\partial m_\mu \partial q = 0$$

The $(\ell{+}1) \times (\ell{+}1)$ matrix of second derivatives with respect to variation of $(\boldsymbol{m}, q)$, evaluated in the spin-glass saddle-point, thereby acquires a diagonal form

$$\partial^2 f = \begin{pmatrix} 1-\beta(1-q) & & & \\ & \ddots & & \\ & & 1-\beta(1-q) & \\ & & & \partial^2 f/\partial q^2 \end{pmatrix}$$

and the eigenvalues can simply be read off. The ℓ-fold degenerate eigenvalue $1-\beta(1-q)$ is always positive (otherwise (2.55) would not even exist), implying stability of the spin-glass state in the direction of the nominated patterns. The remaining eigenvalue measure the stability of the spin-glass state with respect to variation in the amplitude q. Below the critical temperature $T = 1+\sqrt{\alpha}$ it turns out to be positive for the spin-glass solution of (2.57) with nonzero q. One important difference between the previously studied case $\alpha = 0$ and the present case $\alpha > 0$ is that there is now a $\boldsymbol{m} = 0$ spin-glass solution which is *stable* for all $T < 1+\sqrt{\alpha}$. In terms of information processing this implies that for $\alpha > 0$ an initial state must have a certain non-zero overlap with a pattern to evoke a final state with $\boldsymbol{m} \neq 0$, in order to avoid ending up in the $\boldsymbol{m} = 0$ spin-glass state. In contrast, for $\alpha = 0$, the state with $\boldsymbol{m} = 0$ is unstable, so *any* initial state will eventually lead to a final state with $\boldsymbol{m} \neq 0$.

Inserting the pure state ansatz $\boldsymbol{m} = m(1,0,\ldots,0)$ into our RS equations gives

$$m = \int Dz \;\tanh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-\beta(1-q)}\right] \tag{2.59}$$

$$q = \int Dz \;\tanh^2\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-\beta(1-q)}\right] \tag{2.60}$$

$$f = \frac{1}{2}m^2 + \frac{1}{2}\alpha\left[(1-q)\frac{1+\beta(1-q)(\beta-2)}{[1-\beta(1-q)]^2} + \frac{1}{\beta}\log\left[1-\beta(1-q)\right]\right] - \frac{1}{\beta}\int Dz \;\log 2\cosh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-\beta(1-q)}\right] \tag{2.61}$$

If we solve the equations (2.59,2.60) numerically for different values of α, and calculate the corresponding 'free energies' f (2.61) for the pure states and the spin-glass state $\boldsymbol{m} = 0$, we obtain figure 2.7. For $\alpha > 0$ the nontrivial solution m for the amplitude of the pure state appears *discontinously* as the temperature is lowered, defining a critical temperature $T_M(\alpha)$. Once the pure state appears, it turns out to be locally stable (within the RS ansatz). Its 'free energy' f, however, remains larger than the one corresponding to the spin-glass state, until the temperature is further reduced to below a second critical temperature $T_c(\alpha)$. For $T < T_c(\alpha)$ the pure states are therefore the equilibrium states in the thermodynamics sense.

By drawing these critical lines in the (α, T) plane, together with the line $T_g(\alpha) = 1+\sqrt{\alpha}$ which signals the second order transition from the paramagnetic to the spin-glass state, we obtain the RS phase diagram of the Hopfield model, depicted in figure 2.8. Strictly speaking the line T_M would appear meaningless in the thermodynamic picture, only the saddle-point that minimises f being relevant. However, we have to keep in mind the physics behind the formalism. The occurrence of multiple locally stable saddle-points is the manifestation of ergodicity breaking in the limit $N \to \infty$. The thermodynamic analysis, based on ergodicity, therefore applies only within a single ergodic component. As a result, each locally stable saddle-point is indeed relevant for appropriate initial conditions and time-scales.

Zero Temperature, Storage Capacity. The storage capacity α_c of the Hopfield model is defined as the largest α for which locally stable pure states exist. If for the moment we neglect the low temperature re-entrance peculiarities in the phase diagram (2.8) to which we will come back later, the critical temperature $T_M(\alpha)$, where the pure states appear decreases monotonically with α, and the storage capacity is reached for $T = 0$. Before we can put $T \to 0$ in (2.59, 2.60), however, we will have to rewrite these equations in terms of quantities with well defined $T \to 0$ limits, since $q \to 1$. A suitable quantity is $C = \beta(1-q)$, which obeys $0 \leq C \leq 1$ for the free

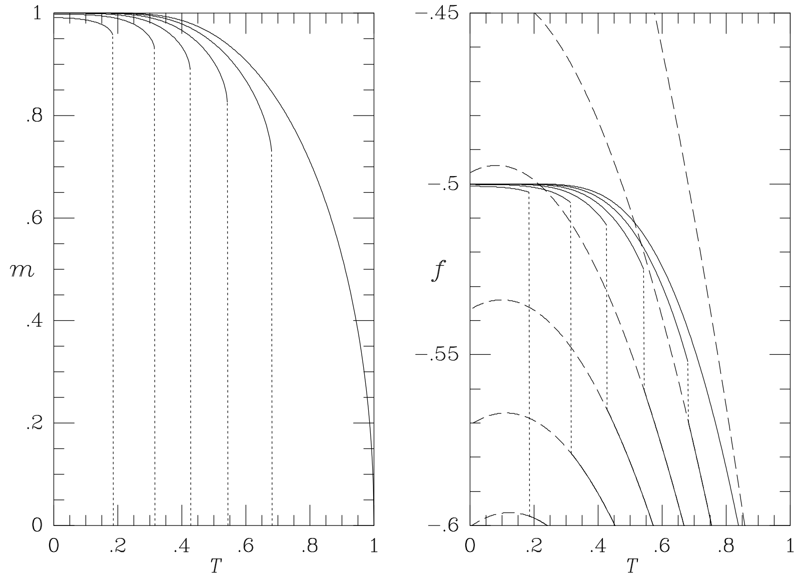


Figure 2.7: Left picture: RS amplitudes m of the pure states of the Hopfield model as a function of temperature. From top to bottom: $\alpha = 0.000 - 0.125$ ($\Delta\alpha = 0.025$). Right picture, solid lines: 'free energies' f of the pure states. From bottom to top: $\alpha = 0.000 - 0.125$ ($\Delta\alpha = 0.025$). Right picture, dashed lines: 'free energies' of the spin-glass state $\boldsymbol{m} = 0$ (for comparison). From top to bottom: $\alpha = 0.000 - 0.125$ ($\Delta\alpha = 0.025$).

energy (2.55) to exist. The saddle-point equations can now be written in the form

$$m = \int Dz\ \tanh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-C}\right] \qquad C = \frac{\partial}{\partial m}\int Dz\ \tanh\left[\beta m + \frac{z\beta\sqrt{\alpha q}}{1-C}\right]$$

in which the limit $T \to 0$ simply corresponds to $\tanh(\beta x) \to \text{sgn}(x)$ and $q \to 1$. After having taken the limit we perform the Gaussian integral:

$$m = \text{erf}\left[\frac{m(1-C)}{\sqrt{2\alpha}}\right] \qquad C = (1-C)\sqrt{\frac{2}{\alpha\pi}}e^{-m^2(1-C)^2/2\alpha}$$

This set can be reduced to a single trancendental equation by introducing $x = m(1-C)/\sqrt{2\alpha}$:

$$x\sqrt{2\alpha} = F(x) \qquad F(x) = \text{erf}(x) - \frac{2x}{\sqrt{\pi}}e^{-x^2} \tag{2.62}$$

Equation (2.62) is solved numerically (see figure 2.9). Since $F(x)$ is anti-symmetric, solutions come in pairs $(x, -x)$ (reflecting the symmetry of the Hamiltonian of the system with respect to an overall spin-flip $\boldsymbol{\sigma} \to -\boldsymbol{\sigma}$). For $\alpha < \alpha_c \sim 0.138$ there indeed exists pure state solutions $x \neq 0$. For $\alpha > \alpha_c$ there is only the spin-glass solution $x = 0$. Given a solution x of (2.62), the zero temperature values for the order parameters follow from

$$\lim_{T\to 0} m = \text{erf}[x] \qquad \lim_{T\to 0} C = \left[1 + \sqrt{\frac{\alpha\pi}{2}}e^{x^2}\right]^{-1}$$

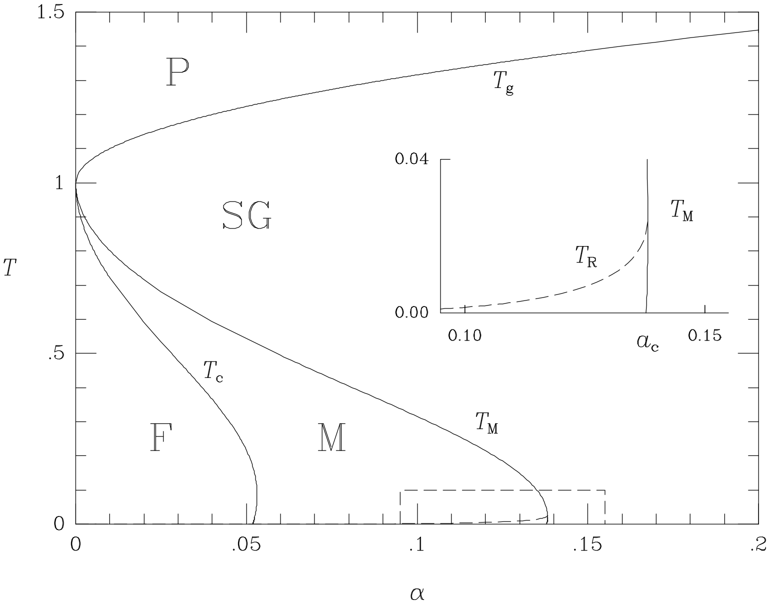


Figure 2.8: Phase diagram of the Hopfield model. P: paramagnetic phase, $m = q = 0$. SG: spin-glass phase, $m = 0$, $q \neq 0$. F: pattern recall phase (the pure states minimise f), $m \neq 0$, $q \neq 0$. M: mixed phase (the pure states are local but not global minima of f). Solid lines: separations of the above phases (T_g: second order, T_M and T_c: first order). Dashed: the AT instability for the retrieval solutions (T_R). Inset: close-up of the low temperature region.

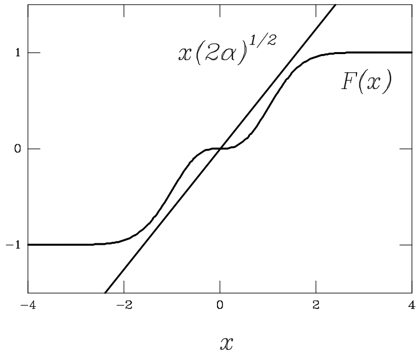


Figure 2.9: Solution of the transcendental equation $F(x) = x\sqrt{2\alpha}$, where $x = \mathrm{erf}^{\mathrm{inv}}(m)$. The storage capacity $\alpha_c \sim 0.138$ of the Hopfield model is the largest α for which solutions $x \neq 0$ exist.

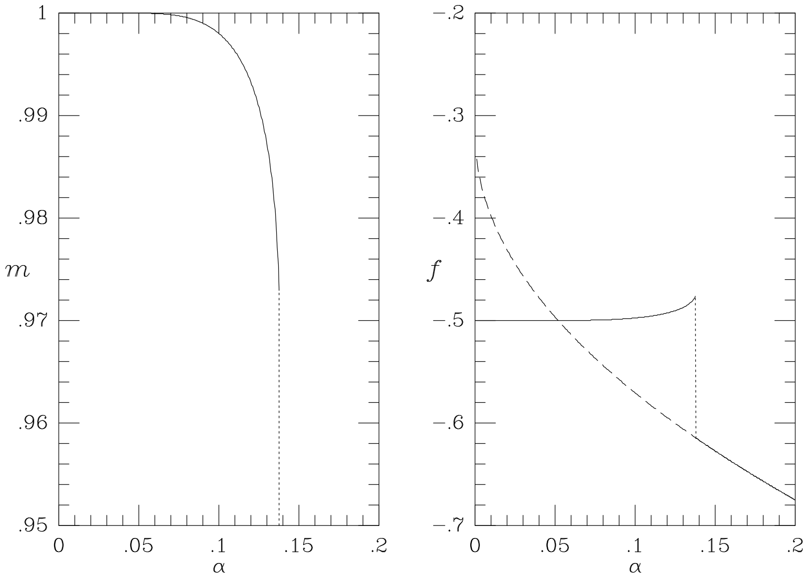


Figure 2.10: Left picture: RS amplitudes m of the pure states of the Hopfield model for $T = 0$ as a function of $\alpha = p/N$. The location of the discontinuity, where m vanishes, defines the storage capacity $\alpha_c \sim 0.138$. Right picture, solid line: $T = 0$ 'free energy' f of the pure states. Dashed lines: $T = 0$ 'free energy' of the spin-glass state $\boldsymbol{m} = 0$ (for comparison).

with which in turn we can take the zero temperature limit in our expression (2.61) for the free energy:

$$\lim_{T\to 0} f = \frac{1}{2}\text{erf}^2[x] + \frac{1}{\pi}e^{-x^2} - \frac{2}{\pi}\left[e^{-x^2} + \sqrt{\frac{\alpha\pi}{2}}\right]\left[x\sqrt{\pi}\ \text{erf}[x] + e^{-x^2}\right]$$

Comparison of the values for $\lim_{T\to 0} f$ thus obtained, for the pure state $m > 0$ and the spin-glass state $m = 0$ leads to figure 2.10, which clearly shows that for sufficiently small α the pure states are the true ground states of the system.

Breaking of Replica Symmetry: The AT Instability. As in the case of the SK spin-glass model the above RS solution generates negative entropies at sufficiently low temperatures, indicating that replica-symmetry must be broken. If saddle-points without replica symmetry bifurcate continuously from the RS one, we can locate the replica symmetry breaking by studying the effect on $f(\boldsymbol{m}, \boldsymbol{q}, \hat{\boldsymbol{q}})$ (2.45) of small replicon fluctuations around the RS solution, á la de Almeida and Thouless (1978):

$$q_{\alpha\beta} \to \delta_{\alpha\beta} + q\,[1 - \delta_{\alpha\beta}] + \eta_{\alpha\beta}$$

$$\eta_{\alpha\beta} = \eta_{\beta\alpha} \qquad \eta_{\alpha\alpha} = 0 \qquad \sum_{\alpha} \eta_{\alpha\beta} = 0 \tag{2.63}$$

The variation of $\boldsymbol{q}$ induces a similar variation in the conjugate parameters $\hat{\boldsymbol{q}}$ through equation

(2.48):

$$\hat{q}_{\alpha\beta} \to \frac{1}{2} i\alpha\beta^2 \left[R\delta_{\alpha\beta} + r\left[1-\delta_{\alpha\beta}\right] + \hat{\eta}_{\alpha\beta}\right]$$

$$\hat{\eta}_{\alpha\beta} = \frac{1}{2}\sum_{\gamma\delta} \eta_{\gamma\delta}\left[g_{\alpha\beta\gamma\delta} - g_{\alpha\beta}g_{\gamma\delta}\right]$$

with

$$g_{\alpha\beta\gamma\delta} = \frac{\int d\boldsymbol{z}\; z_\alpha z_\beta z_\gamma z_\delta e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbf{I}-\beta\boldsymbol{q}_{\rm RS}]\boldsymbol{z}}}{\int d\boldsymbol{z}\; e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbf{I}-\beta\boldsymbol{q}_{\rm RS}]\boldsymbol{z}}} \qquad\qquad g_{\alpha\beta} = \frac{\int d\boldsymbol{z}\; z_\alpha z_\beta e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbf{I}-\beta\boldsymbol{q}_{\rm RS}]\boldsymbol{z}}}{\int d\boldsymbol{z}\; e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbf{I}-\beta\boldsymbol{q}_{\rm RS}]\boldsymbol{z}}}$$

Wick's theorem (see e.g. Zinn-Justin (1993)) can now be used to write everything in terms of second moments of the Gaussian integrals only:

$$g_{\alpha\beta\gamma\delta} = g_{\alpha\beta}g_{\gamma\delta} + g_{\alpha\gamma}g_{\beta\delta} + g_{\alpha\delta}g_{\beta\gamma}$$

with which we can express the replicon variation in $\hat{\boldsymbol{q}}$, using the symmetry of $\{\eta_{\alpha\beta}\}$ and the saddle-point equation (2.48), as

$$\begin{aligned} \hat{\eta}_{\alpha\beta} &= \textstyle\sum_{\gamma\delta} g_{\alpha\gamma}\eta_{\gamma\delta}g_{\delta\beta} \\ &= \beta^2 \textstyle\sum_{\gamma\neq\delta}\left[R\delta_{\alpha\gamma}+r\left[1-\delta_{\alpha\gamma}\right]\right]\eta_{\gamma\delta}\left[R\delta_{\delta\beta}+r\left[1-\delta_{\delta\beta}\right]\right] \\ &= \beta^2(R-r)^2\eta_{\alpha\beta} \end{aligned} \tag{2.64}$$

since only those terms can contribute which involve precisely two δ-symbols, as a consequence of $\sum_\alpha \eta_{\alpha\beta} = 0$.

We can now calculate the resulting change in $f(\boldsymbol{m},\boldsymbol{q},\hat{\boldsymbol{q}})$, away from the RS value $f(\boldsymbol{m}_{\rm RS},\boldsymbol{q}_{\rm RS},\hat{\boldsymbol{q}}_{\rm RS})$, the leading order of which must be quadratic in the fluctuations $\{\eta_{\alpha\beta}\}$ since the RS solution (2.56,2.57,2.58) is a saddle-point:

$$\begin{aligned} f(\boldsymbol{m}_{\rm RS},\boldsymbol{q},\hat{\boldsymbol{q}}) - f(\boldsymbol{m}_{\rm RS},\boldsymbol{q}_{\rm RS},\hat{\boldsymbol{q}}_{\rm RS}) &= \frac{1}{\beta n}\left[\frac{1}{2}\alpha\log\frac{\det\left[\mathbf{I}-\beta(\boldsymbol{q}_{\rm RS}+\boldsymbol{\eta})\right]}{\det\left[\mathbf{I}-\beta\boldsymbol{q}_{\rm RS}\right]} - i{\rm Tr}\left[\hat{\boldsymbol{q}}_{\rm RS}.\boldsymbol{\eta}\right]\right. \\ &\left. +\frac{1}{2}\alpha\beta^2{\rm Tr}\left[\hat{\boldsymbol{\eta}}.\boldsymbol{\eta}+\hat{\boldsymbol{\eta}}.\boldsymbol{q}_{\rm RS}\right] - \langle\log\frac{\langle e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\rm RS}\sum_\alpha\sigma_\alpha - i\boldsymbol{\sigma}\cdot[\hat{\boldsymbol{q}}_{\rm RS}+\frac{1}{2}i\alpha\beta^2\hat{\boldsymbol{\eta}}]\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\rm RS}\sum_\alpha\sigma_\alpha - i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\rm RS}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\rangle_{\boldsymbol{\xi}}\right] \end{aligned} \tag{2.65}$$

This expression looks more awkward than it actually is. Evaluation is greatly simplified by the fact that the matrices $\boldsymbol{q}_{\rm RS}$ and $\boldsymbol{\eta}$ commute, which is a direct consequence of the properties (2.63) of the replicon fluctuations and the form of the replica-symmetric saddle-point. If we define the $n\times n$ matrix $\boldsymbol{P}$ to be the projection onto the uniform state $(1,\ldots,1)$, we have the relations

$$\begin{gathered} P_{\alpha\beta} = \frac{1}{n} \qquad \boldsymbol{P}.\boldsymbol{\eta} = \boldsymbol{\eta}.\boldsymbol{P} = 0 \qquad \boldsymbol{q}_{\rm RS} = (1-q)\mathbf{I} + nq\boldsymbol{P} \\ \boldsymbol{q}_{\rm RS}.\boldsymbol{\eta} = \boldsymbol{\eta}.\boldsymbol{q}_{\rm RS} = (1-q)\boldsymbol{\eta} \\ \left[\mathbf{I}-\beta\boldsymbol{q}_{\rm RS}\right]^{-1} = \frac{1}{1-\beta(1-q)}\mathbf{I} + \frac{\beta nq}{\left[1-\beta(1-q)-\beta nq\right]\left[1-\beta(1-q)\right]}\boldsymbol{P} \end{gathered} \tag{2.66}$$

We can now simply expand the relevant terms, using the identity $\log\det M = \mathrm{Tr}\ \log M$:

$$\log\frac{\det\left[\mathbf{I}-\beta(\boldsymbol{q}_{\mathrm{RS}}+\boldsymbol{\eta})\right]}{\det\left[\mathbf{I}-\beta\boldsymbol{q}_{\mathrm{RS}}\right]} = \mathrm{Tr}\log\left[\mathbf{I}-\beta\boldsymbol{\eta}\left[\mathbf{I}-\beta\boldsymbol{q}_{\mathrm{RS}}\right]^{-1}\right]$$

$$= \mathrm{Tr}\left\{-\beta\boldsymbol{\eta}\left[\mathbf{I}-\beta\boldsymbol{q}_{\mathrm{RS}}\right]^{-1}-\frac{1}{2}\beta^2\left[\boldsymbol{\eta}\left[\mathbf{I}-\beta\boldsymbol{q}_{\mathrm{RS}}\right]^{-1}\right]^2\right\}+\mathcal{O}(\boldsymbol{\eta}^3)$$

$$= -\frac{1}{2}\frac{\beta^2}{[1-\beta(1-q)]^2}\mathrm{Tr}\ \boldsymbol{\eta}^2+\mathcal{O}(\boldsymbol{\eta}^3) \tag{2.67}$$

Finally we address the remaining term in (2.65), again using the RS saddle-point equations (2.56,2.57,2.58) where appropriate:

$$\langle\log\frac{\langle e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\left[1+\frac{1}{2}\alpha\beta^2\boldsymbol{\sigma}\cdot\hat{\boldsymbol{\eta}}\boldsymbol{\sigma}+\frac{1}{8}\alpha^2\beta^4(\boldsymbol{\sigma}\cdot\hat{\boldsymbol{\eta}}\boldsymbol{\sigma})^2+\ldots\right]\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\rangle_{\boldsymbol{\xi}}$$

$$= \frac{1}{2}\alpha\beta^2\mathrm{Tr}[\hat{\boldsymbol{\eta}}.\boldsymbol{q}_{\mathrm{RS}}]+\frac{1}{8}\alpha^2\beta^4\sum_{\alpha\beta\gamma\delta}\hat{\eta}_{\alpha\beta}\hat{\eta}_{\gamma\delta}[G_{\alpha\beta\gamma\delta}-H_{\alpha\beta\gamma\delta}]+\ldots \tag{2.68}$$

with

$$G_{\alpha\beta\gamma\delta} = \langle\frac{\langle\sigma_\alpha\sigma_\beta\sigma_\gamma\sigma_\delta e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\rangle_{\boldsymbol{\xi}}$$

$$H_{\alpha\beta\gamma\delta} = \langle\frac{\langle\sigma_\alpha\sigma_\beta e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\frac{\langle\sigma_\gamma\sigma_\delta e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\boldsymbol{\xi}\cdot\boldsymbol{m}_{\mathrm{RS}}\sum_\alpha\sigma_\alpha-i\boldsymbol{\sigma}\cdot\hat{\boldsymbol{q}}_{\mathrm{RS}}\boldsymbol{\sigma}}\rangle_{\boldsymbol{\sigma}}}\rangle_{\boldsymbol{\xi}}$$

Inserting the ingredients (2.64,2.66,2.67,2.68) into expression (2.65) and rearranging terms shows that the linear terms indeed cancel and that the term involving $H_{\alpha\beta\gamma\delta}$ does not contribute (since the elements $H_{\alpha\beta\gamma\delta}$ don't depend on the indices for $\alpha\neq\beta$ and $\gamma\neq\delta$), and we are left with:

$$f(\boldsymbol{m}_{\mathrm{RS}},\boldsymbol{q},\hat{\boldsymbol{q}})-f(\boldsymbol{m}_{\mathrm{RS}},\boldsymbol{q}_{\mathrm{RS}},\hat{\boldsymbol{q}}_{\mathrm{RS}}) = \frac{1}{\beta n}\left[-\frac{1}{4}\frac{\alpha\beta^2}{[1-\beta(1-q)]^2}\mathrm{Tr}\ \boldsymbol{\eta}^2+\frac{1}{2}\alpha\beta^4(R-r)^2\mathrm{Tr}\ \boldsymbol{\eta}^2\right.$$

$$\left.-\frac{1}{8}\alpha^2\beta^8(R-r)^4\sum_{\alpha\beta\gamma\delta}\eta_{\alpha\beta}\eta_{\gamma\delta}G_{\alpha\beta\gamma\delta}\right]+\ldots$$

Because of the index permutation symmetry in the spin-average we can write for $\alpha\neq\gamma$ and $\rho\neq\lambda$:

$$G_{\alpha\gamma\rho\lambda} = \delta_{\alpha\rho}\delta_{\gamma\lambda}+\delta_{\alpha\lambda}\delta_{\gamma\rho}+G_4\,[1-\delta_{\alpha\rho}]\,[1-\delta_{\gamma\lambda}]\,[1-\delta_{\alpha\lambda}]\,[1-\delta_{\gamma\rho}]$$

$$+\ G_2\,\{\delta_{\alpha\rho}\,[1-\delta_{\gamma\lambda}]+\delta_{\gamma\lambda}\,[1-\delta_{\alpha\rho}]+\delta_{\alpha\lambda}\,[1-\delta_{\gamma\rho}]+\delta_{\gamma\rho}\,[1-\delta_{\alpha\lambda}]\}$$

with

$$G_\ell = \langle\frac{\int Dz\ \tanh^\ell\beta\,[\boldsymbol{m}\cdot\boldsymbol{\xi}+z\sqrt{\alpha r}]\cosh^n\beta\,[\boldsymbol{m}\cdot\boldsymbol{\xi}+z\sqrt{\alpha r}]}{\int Dz\ \cosh^n\beta\,[\boldsymbol{m}\cdot\boldsymbol{\xi}+z\sqrt{\alpha r}]}\rangle_{\boldsymbol{\xi}}$$

Only terms which involve precisely two δ-functions can contribute, because of the replicon properties (2.63). As a result:

$$f(\boldsymbol{m}_{\rm RS},\boldsymbol{q},\hat{\boldsymbol{q}})-f(\boldsymbol{m}_{\rm RS},\boldsymbol{q}_{\rm RS},\hat{\boldsymbol{q}}_{\rm RS})=\frac{1}{\beta n}{\rm Tr}\ \boldsymbol{\eta}^2\left[-\frac{1}{4}\frac{\alpha\beta^2}{[1-\beta(1-q)]^2}+\frac{1}{2}\alpha\beta^4(R-r)^2\right.$$

$$\left.-\frac{1}{4}\alpha^2\beta^8(R-r)^4\,[1-2G_2+G_4]\right]+\ldots$$

Since ${\rm Tr}\ \boldsymbol{\eta}^2=\sum_{\alpha\beta}\eta_{\alpha\beta}^2$, the condition for the RS solution to minimise $f(\boldsymbol{m},\boldsymbol{q},\hat{\boldsymbol{q}})$, if compared to the 'replicon' fluctuations, is therefore

$$-\frac{1}{[1-\beta(1-q)]^2}+2\beta^2(R-r)^2-\alpha\beta^6(R-r)^4\,[1-2G_2+G_4]>0 \tag{2.69}$$

After taking the limit in the expressions G_ℓ and after evaluating

$$\lim_{n\to 0}R=\frac{1}{\beta}\lim_{n\to 0}g_{\alpha\alpha}=\lim_{n\to 0}\frac{1}{n\beta}\frac{\int\!d\boldsymbol{z}\ \boldsymbol{z}^2e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbb{1}-\beta\boldsymbol{q}_{\rm RS}]\boldsymbol{z}}}{\int\!d\boldsymbol{z}\ e^{-\frac{1}{2}\boldsymbol{z}\cdot[\mathbb{1}-\beta\boldsymbol{q}_{\rm RS}]\boldsymbol{z}}}$$

$$=\lim_{n\to 0}\frac{1}{n\beta}\left[\frac{n-1}{1-\beta(1-q)}+\frac{1}{1-\beta(1-q+nq)}\right]=\frac{1}{\beta}\frac{1-\beta+2\beta q}{[1-\beta(1-q)]^2}$$

and using (2.58), the condition (2.69) can be written as

$$1>\frac{\alpha\beta^2}{[1-\beta(1-q)]^2}\langle\int\!Dz\ \cosh^{-4}\beta\left[\boldsymbol{m}\cdot\boldsymbol{\xi}+z\sqrt{\alpha r}\right]\rangle_{\boldsymbol{\xi}} \tag{2.70}$$

The AT line in the phase diagram, where this condition ceases to be met, indicates a second-order transition to a spin-glass state where ergodicity is broken in the sense that the distribution $\overline{P(q)}$ (2.51) is no longer a δ-function.

In the paramagnetic regime of the phase diagram, $\boldsymbol{m}=q=0$, the AT condition reduces precisely to $T>T_g=1+\sqrt{\alpha}$. Therefore the paramagnetic solution is stable. The AT line coincides with the boundary between the paramagnetic and spin-glass phase. Numerical evaluation of (2.70) shows that the RS spin-glass solution remains unstable for all $T<T_g$, but that the retrieval solution $\boldsymbol{m}\neq 0$ is unstable only for very low temperatures $T<T_R$ (see figure 2.8).

Chapter 3

Dynamics Away from Saturation

Techniques from equilibrium statistical mechanics can provide much detailed quantitative information on the behaviour of large neural networks, but they also have some serious restrictions. The first (obvious) one is that, by definition, they will only provide information on equilibrium properties. For associative memories, for instance, it is not clear how one can calculate quantities like sizes of domains of attraction without studying dynamics. The second (more serious) restriction is that for equilibrium statistical mechanics to apply the dynamics of the system under study must obey detailed balance. As we have seen, for Ising spin neural networks in which the dynamics is a stochastic alignment to local fields (or post-synaptic potentials) which are linear in the neural state variables, as in (1.1,1.5), this requirement implies immediately symmetry of the interaction matrix. From a physiological point of view this is unacceptable. For non-symmetric systems, ergodicity breaking in the thermodynamic limit (i.e. on finite timescales) may now manifest itself in the form of limit-cycle attractors or even in chaotic trajectories, rather than just fixed-point attractors.

microscopic level:
N Ising spins σ_i

$$\begin{tikzcd}
\text{eqn for } p_t(\boldsymbol{\sigma}) \arrow[r, "\textit{solve}"] \arrow[d, "\textit{count states}"'] \arrow[dr, phantom, "P_t(\mathbf{\Omega})=\sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})\delta\left[\mathbf{\Omega}-\mathbf{\Omega}(\boldsymbol{\sigma})\right]"] & p_t(\boldsymbol{\sigma})=\ldots \arrow[d, "\textit{count states}"] \\
\text{eqn. for } P_t(\mathbf{\Omega}) \arrow[r, "\textit{solve}"'] & P_t(\mathbf{\Omega})=\ldots
\end{tikzcd}$$

macroscopic level:
ℓ order parameters $\Omega_\mu(\boldsymbol{\sigma})$

The common strategy of all non-equilibrium statistical mechanical studies is to derive and solve dynamical laws for a suitable smaller set of relevant macroscopic quantities $\mathbf{\Omega}$ from the dynamical laws of the underlying microscopic system $\boldsymbol{\sigma}$. This can be done in two ways (see diagram). The first route consists of calculating from the microscopic stochastic equations an equation for the macroscopic probability distribution, which subsequently is to be solved. If applicable, this is the easiest one. The second route consists of solving the microscopic stochastic equations directly; from this solution one then calculates the values of the macroscopic quantities. Away from saturation we can usually formulate and solve the problem in terms of true state variables, corresponding to macroscopic quantities that can be measured instantly and therefore carry only one time-argument (like, for instance, magnetisation and energy). Near saturation this will be no longer feasable, and we must introduce quantities with two time-arguments: correlation- and response functions.

For such programmes to work the interaction matrix must either have a suitable structure of some sort, or contain (frozen) disorder, over which suitable averages can be performed (or a combination of both). A common feature of many statistical mechanical models for neural networks is separability of the interaction matrix, which naturally leads to a convenient description in terms of macroscopic order parameters.

3.1 Sequential Dynamics

In this section we show how for sequential dynamics one can calculate from the microscopic stochastic evolution equations (at the level of individual neurons) differential equations for the probability distribution of suitably defined macroscopic state variables. For mathematical convenience our starting point will be the continuous-time master equation (1.12), rather than the discrete version (1.5). We will investigate which are the conditions for the evolution of these macroscopic state variables to become deterministic in the limit of infinitely large networks and, in addition, be governed by a closed set of dynamic equations. Finally we turn to certain classes of models, with and without detailed balance, and show how the macroscopic equations can be used to illuminate and understand the dynamics of attractor neural networks away from saturation.

A Toy Model. Let us first illustrate the basic ideas with the help of a simple toy model:

$$J_{ij} \equiv \frac{J}{N}\eta_i\xi_j \qquad\qquad \theta_i \equiv 0 \tag{3.1}$$

(the variables η_i and ξ_i are arbitrary, but may not depend on N). For $\eta_i = \xi_i = 1$ we recover the infinite range ferromagnet ($J > 0$) or anti-ferromagnet ($J < 0$); for $\eta_i = \xi_i \in \{-1, 1\}$ (random) and $J > 0$ we recover the Lüttinger or Mattis model (equivalently: the Hopfield model with only one stored pattern). Note, however, that the interaction matrix is non-symmetric as soon as a pair (ij) exists, such that $\eta_i\xi_j \neq \eta_j\xi_i$ (in general, therefore, equilibrium statistical mechanics does not apply). The local fields become $h_i(\boldsymbol{\sigma}) = J\eta_i m(\boldsymbol{\sigma})$ with $m(\boldsymbol{\sigma}) \equiv \frac{1}{N}\sum_k \xi_k\sigma_k$. Since they depend on the microscopic state $\boldsymbol{\sigma}$ only through the value of m, the latter quantity appears to constitute a natural macroscopic level of description. The ensemble probability of finding the macroscopic state $m(\boldsymbol{\sigma}) = m$ is given by

$$\mathcal{P}_t[m] \equiv \sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})\delta\left[m - m(\boldsymbol{\sigma})\right]$$

Its time derivative is obtained by inserting (1.12):

$$\frac{d}{dt}\mathcal{P}_t[m] = \sum_{\boldsymbol{\sigma}}\sum_{k=1}^{N} p_t(\boldsymbol{\sigma})w_k(\boldsymbol{\sigma})\left\{\delta\left[m - m(\boldsymbol{\sigma}) + \frac{2}{N}\xi_k\sigma_k\right] - \delta\left[m - m(\boldsymbol{\sigma})\right]\right\}$$

$$= \frac{d}{dm}\left\{\sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})\delta\left[m - m(\boldsymbol{\sigma})\right]\frac{2}{N}\sum_{k=1}^{N}\xi_k\sigma_k w_k(\boldsymbol{\sigma})\right\} + \mathcal{O}(\frac{1}{N})$$

Inserting the expression (1.9) for the transition rates and the local fields gives:

$$\frac{d}{dt}\mathcal{P}_t[m] = \frac{d}{dm}\left\{\mathcal{P}_t[m]\left[m - \frac{1}{N}\sum_{k=1}^{N}\xi_k \tanh[\eta_k\beta J m]\right]\right\} + \mathcal{O}(N^{-1})$$

In the thermodynamic limit $N \to \infty$ only the first term survives. The solution of the resulting differential equation for $\mathcal{P}_t[m]$ is:

$$\mathcal{P}_t[m] = \int dm_0\ \mathcal{P}_0[m_0]\delta\left[m - m^*(t)\right]$$

$$\frac{d}{dt}m^* = \lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\xi_k \tanh[\eta_k\beta J m^*] - m^* \qquad m^*(0) \equiv m_0 \tag{3.2}$$

This solution describes deterministic evolution, the only uncertainty in the value of m is due to uncertainty in initial conditions. If at $t = 0$ the quantity m is known exactly, this will remain the case for finite timescales; m turns out to evolve in time according to (3.2).

Arbitrary Synaptic Interactions. Let us now allow for less trivial choices of the interaction matrix and try to calculate the evolution in time of a given set of macroscopic state variables $\boldsymbol{\Omega}(\boldsymbol{\sigma}) \equiv (\Omega_1(\boldsymbol{\sigma}), \ldots, \Omega_n(\boldsymbol{\sigma}))$ in the thermodynamic limit $N \to \infty$. At this stage there are no restrictions yet on the form or the number n of these state variables; such conditions, however, naturally arise if we require the evolution of the variables $\boldsymbol{\Omega}$ to obey a closed set of deterministic laws, as we will show below. The ensemble probability of finding the system in macroscopic state $\boldsymbol{\Omega}$ is given by:

$$\mathcal{P}_t\left[\boldsymbol{\Omega}\right] \equiv \sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})\delta\left[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{\sigma})\right]$$

The time derivative of this distribution is obtained by inserting (1.12). If in those parts of the resulting expression which contain the spin-flip operators F_i we subsequently perform gauge transformations $\boldsymbol{\sigma} \to F_i\boldsymbol{\sigma}$, we arrive at

$$\frac{d}{dt}\mathcal{P}_t\left[\boldsymbol{\Omega}\right] = \sum_i\sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})w_i(\boldsymbol{\sigma})\left\{\delta\left[\boldsymbol{\Omega} - \boldsymbol{\Omega}(F_i\boldsymbol{\sigma})\right] - \delta\left[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{\sigma})\right]\right\}$$

Upon writing $\Omega_\mu(F_i\boldsymbol{\sigma}) = \Omega_\mu(\boldsymbol{\sigma}) + \Delta_{i\mu}(\boldsymbol{\sigma})$ and making a Taylor expansion in powers of $\{\Delta_{i\mu}(\boldsymbol{\sigma})\}$, we finally obtain the so-called Kramers-Moyal expansion:

$$\frac{d}{dt}\mathcal{P}_t\left[\boldsymbol{\Omega}\right] = \sum_{l\geq 1}\frac{(-1)^l}{l!}\sum_{\mu_1=1}^{n}\cdots\sum_{\mu_l=1}^{n}\frac{\partial^l}{\partial\Omega_{\mu_1}\cdots\partial\Omega_{\mu_l}}\left\{\mathcal{P}_t\left[\boldsymbol{\Omega}\right]F^{(l)}_{\mu_1\cdots\mu_l}\left[\boldsymbol{\Omega};t\right]\right\} \tag{3.3}$$

which is defined in terms of sub-shell averages $\langle f(\boldsymbol{\sigma})\rangle_{\boldsymbol{\Omega};t}$ and the 'discrete derivatives' $\Delta_{j\mu}(\boldsymbol{\sigma}) = \Omega_\mu(F_j\boldsymbol{\sigma}) - \Omega_\mu(\boldsymbol{\sigma})$:

$$F^{(l)}_{\mu_1\cdots\mu_l}[\boldsymbol{\Omega};t] = \langle\sum_{j=1}^{N} w_j(\boldsymbol{\sigma})\Delta_{j\mu_1}(\boldsymbol{\sigma})\cdots\Delta_{j\mu_l}(\boldsymbol{\sigma})\rangle_{\boldsymbol{\omega};t}$$

$$\langle f(\boldsymbol{\sigma})\rangle_{\boldsymbol{\Omega};t} = \frac{\sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})\delta\left[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{\sigma})\right] f(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})\delta\left[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{\sigma})\right]}$$

The expansion (3.3) is to be interpreted in a distributional sense, i.e. only to be used in expressions of the form $\int d\boldsymbol{\Omega}\mathcal{P}_t(\boldsymbol{\Omega})G(\boldsymbol{\Omega})$ with sufficiently smooth functions $G(\boldsymbol{\Omega})$, so that all derivatives are well-defined and finite. Furthermore, (3.3) will only make sense if the discrete derivatives $\Delta_{j\mu}$, which measure the sensitivity of the macroscopic quantities to single spin flips, are sufficiently small. This is to be expected from a physical point of view: for finite N *any* state variable $\Omega_\mu(\boldsymbol{\sigma})$ can only assume a finite number of possible values; only in the limit $N\to\infty$ may we expect smooth probability distributions for our macroscopic quantities (the probability distribution of state variables which only depend on a *small* number of spins, however, will *not* become smooth, whatever the system size). The first ($l=1$) term in the series (3.3) is the flow term; retaining only this term leads us to a Liouville equation which describes deterministic flow in $\boldsymbol{\Omega}$ space, driven by the flow field $\boldsymbol{F}^{(1)}$. Including the second ($l=2$) term as well leads us to a Fokker-Planck equation which (in addition to the flow) describes diffusion of the macroscopic probability density $\mathcal{P}_t[\boldsymbol{\Omega}]$, generated by the diffusion matrix $F^{(2)}_{\mu\nu}$. Note, however, that in general (3.3) need not necessarily constitute a systematic expansion in terms of some small parameter.

According to (3.3) a sufficient condition for the set $\boldsymbol{\Omega}(\boldsymbol{\sigma})$ to evolve in time deterministically in the limit $N\to\infty$ is:

$$\lim_{N\to\infty}\sum_{l\geq 2}\frac{1}{l!}\sum_{\mu_1=1}^{n}\cdots\sum_{\mu_l=1}^{n}\sum_{j=1}^{N}\langle|\Delta_{j\mu_1}(\boldsymbol{\sigma})\cdots\Delta_{j\mu_l}(\boldsymbol{\sigma})|\rangle_{\boldsymbol{\Omega};t} = 0 \tag{3.4}$$

(since now for $N\to\infty$ only the $l=1$ term in (3.3) is retained). In the simple case where the state variables Ω_μ are scale similarly in the sense that all 'derivatives' $\Delta_{j\mu}$ are of the same order in the system size N (i.e. there is a monotonic function $\tilde{\Delta}_N$ such that $\Delta_{j\mu} = \mathcal{O}(\tilde{\Delta}_N)$ for all $j\mu$), for instance, the above criterion becomes:

$$\lim_{N\to\infty} n\tilde{\Delta}_N\sqrt{N} = 0 \tag{3.5}$$

If for a given set of macroscopic quantities the condition (3.4) is satisfied we can for large N describe the evolution of the macroscopic probability density by the Liouville equation:

$$\frac{d}{dt}\mathcal{P}_t[\boldsymbol{\Omega}] = -\sum_{\mu=1}^{n}\frac{\partial}{\partial\Omega_\mu}\left\{\mathcal{P}_t[\boldsymbol{\Omega}]\,F^{(1)}_\mu[\boldsymbol{\Omega};t]\right\}$$

the solution of which describes deterministic flow:

$$\mathcal{P}_t[\boldsymbol{\Omega}] = \int d\boldsymbol{\Omega}_0\mathcal{P}_0[\boldsymbol{\Omega}_0]\,\delta\left[\boldsymbol{\Omega}-\boldsymbol{\Omega}^*(t)\right]$$

$$\frac{d}{dt}\boldsymbol{\Omega}^*(t) = \boldsymbol{F}^{(1)}\left[\boldsymbol{\Omega}^*(t);t\right] \qquad \boldsymbol{\Omega}^*(0) = \boldsymbol{\Omega}_0 \tag{3.6}$$

In taking the limit $N \to \infty$, however, we have to keep in mind that the resulting deterministic theory is obtained by taking this limit for *finite t*. According to (3.3) the $l > 1$ terms do come into play for sufficiently large times t; for $N \to \infty$, however, these times diverge by virtue of (3.4).

Equation (3.6) will in general not be autonomous; tracing back the origin of the explicit time dependence in the right-hand side of (3.6) one finds that in order to calculate $\boldsymbol{F}^{(1)}$ one needs to know the microscopic probability density $p_t(\boldsymbol{\sigma})$. This, in turn, requires solving the master equation (1.12) (which is exactly what one tries to avoid). However, there are elegant ways of avoiding this pitfall. We will discuss two constructions that allow for the elimination of the explicit time dependence in the right-hand side of (3.6) and thereby turn the state variables $\boldsymbol{\Omega}$ and their dynamic equations (3.6) into an autonomous level of description.

The first way out is to choose the macroscopic state variables $\boldsymbol{\omega}$ in such a way that there is no explicit time dependence in the flow field $\boldsymbol{F}^{(1)}[\boldsymbol{\Omega};t]$ (if possible). According to the definition of the flow field this implies making sure that there exists a vector field $\boldsymbol{\Phi}[\boldsymbol{\Omega}]$ such that

$$\lim_{N\to\infty}\sum_{j=1}^{N} w_j(\boldsymbol{\sigma})\boldsymbol{\Delta}_j(\boldsymbol{\sigma}) = \boldsymbol{\Phi}\left[\boldsymbol{\Omega}(\boldsymbol{\sigma})\right] \tag{3.7}$$

(with $\boldsymbol{\Delta}_j \equiv (\Delta_{j1},\ldots,\Delta_{jn})$) in which case the time dependence of $\boldsymbol{F}^{(1)}$ drops out and the macroscopic state vector evolves in time according to:

$$\frac{d}{dt}\boldsymbol{\Omega} = \boldsymbol{\Phi}\left[\boldsymbol{\Omega}\right]$$

The advantage is that no restrictions need to be imposed on the initial microscopic configuration; the disadvantage is that for the method to apply, a suitable separable structure of the interaction matrix is required. If, for instance, the macroscopic state variables Ω_μ depend linearly on the microscopic state variables $\boldsymbol{\sigma}$ (i.e. $\boldsymbol{\Omega}(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\omega}_j\sigma_j$), we obtain (with the transition rates (1.9)):

$$\lim_{N\to\infty}\sum_{j=1}^{N} w_j(\boldsymbol{\sigma})\boldsymbol{\Delta}_j(\boldsymbol{\sigma}) = \lim_{N\to\infty}\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{\omega}_j\tanh(\beta h_j(\boldsymbol{\sigma})) - \boldsymbol{\Omega}$$

in which case it turns out that the only further condition necessary for (3.7) to hold is that all local fields h_k must (in leading order in N) depend on the microscopic state $\boldsymbol{\sigma}$ only through the values of the macroscopic state variables $\boldsymbol{\Omega}$ (since the local fields depend linearly on $\boldsymbol{\sigma}$ this, in turn, implies that the interaction matrix must be separable).

If it is not possible to find a set of macroscopic state variables that satisfies both conditions (3.4,3.7), additional assumptions or restrictions are needed. One natural assumption that allows us to close the hierarchy of dynamical equations and obtain an autonomous flow for the state variables $\boldsymbol{\Omega}$ is to assume equipartitioning of probability in the $\boldsymbol{\Omega}$-subshells of the ensemble, which allows us to make the replacement:

$$\boldsymbol{F}^{(1)}\left[\boldsymbol{\Omega};t\right] \quad\to\quad \boldsymbol{F}^{\mathrm{equi}}\left[\boldsymbol{\Omega}\right] = \frac{\sum_{\boldsymbol{\sigma}}\delta\left[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{\sigma})\right]\sum_j w_j(\boldsymbol{\sigma})\boldsymbol{\Delta}_j(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}}\delta\left[\boldsymbol{\Omega}-\boldsymbol{\Omega}(\boldsymbol{\sigma})\right]}$$

Whether or not the above way of closing the set of equations is allowed will depend on the extent to which the relevant stochastic vector $\sum_{j=1}^{N} w_j(\boldsymbol{\sigma})\boldsymbol{\Delta}_j(\boldsymbol{\sigma})$ is homogenous within the $\boldsymbol{\Omega}$-subshells of the ensemble. At $t=0$ there is no problem, since one can always *choose* the initial microscopic distribution $p_0(\boldsymbol{\sigma})$ to obey equipartioning. In the case of extremely diluted networks this situation is subsequently maintained by assuring that, due to the extreme dilution, no correlations can build up in finite time and equipartitioning will be sustained (we will discuss this approach in more detailed in a subsequent chapter). The advantage of extreme dilution, to which we will come back in more detail in a subsequent section, is that less strict requirements on the structure of the interaction matrix are involved. The disadvantage is that the required sparseness of the interactions (compared to the system size) does not correspond to biological reality.

Next we will show how the above formalism can be applied to networks for which the matrix of interactions J_{ij} has a separable form (which includes most symmetric and non-symmetric Hebbian type attractor models). We will restrict ourselves to models with $\theta_i = 0$; the introduction of non-zero thresholds is straightforward and does not pose new problems.

Separable models: Description at the Level of Sublattice Magnetisations. We consider the following class of models, in which the interaction matrix have the form

$$J_{ij} \equiv \frac{1}{N} Q\left(\boldsymbol{\xi}_i ; \boldsymbol{\xi}_j\right) \qquad \boldsymbol{\xi}_i \equiv (\xi_i^1, \ldots, \xi_i^p) \tag{3.8}$$

The components ξ_i^μ are assumed to be drawn from a finite discrete set Λ which contains n_Λ elements (again the variables ξ_i^μ are not allowed to depend on N). As usual they represent the information ('patterns') to be stored or processed. The Hopfield model, for instance, corresponds to choosing $Q(\boldsymbol{x};\boldsymbol{y}) \equiv \boldsymbol{x}\cdot\boldsymbol{y}$ and $\Lambda \equiv \{-1, 1\}$. One can now introduce a partition of the system $\{1, \ldots, N\}$ into n_Λ^p so-called sublattices $I_{\boldsymbol{\eta}}$:

$$I_{\boldsymbol{\eta}} \equiv \{i |\ \boldsymbol{\xi}_i = \boldsymbol{\eta}\} \qquad \{1, \ldots, N\} \equiv \bigcup_{\boldsymbol{\eta}} I_{\boldsymbol{\eta}} \qquad \boldsymbol{\eta} \in \Lambda^p$$

The number of spins in sublattice $I_{\boldsymbol{\eta}}$ will be denoted by $|I_{\boldsymbol{\eta}}|$ (this number will have to be large). If we choose as our macroscopic state variables the magnetisations within these sublattices, we are able to express the alignment fields h_k solely in terms of macroscopic quantities:

$$m_{\boldsymbol{\eta}}(\boldsymbol{\sigma}) \equiv \frac{1}{|I_{\boldsymbol{\eta}}|} \sum_{i \in I_{\boldsymbol{\eta}}} \sigma_i \qquad h_k(\boldsymbol{\sigma}) = \sum_{\boldsymbol{\eta}} p_{\boldsymbol{\eta}} Q\left(\boldsymbol{\xi}_k ; \boldsymbol{\eta}\right) m_{\boldsymbol{\eta}} \tag{3.9}$$

with the relative sublattice sizes $p_{\boldsymbol{\eta}} \equiv |I_{\boldsymbol{\eta}}|/N$. If all $p_{\boldsymbol{\eta}}$ are of the same order in N (which, for example, is the case if the vectors $\boldsymbol{\xi}_i$ have been drawn at random from the set Λ^p) we may write $\Delta_{j\boldsymbol{\eta}} = \mathcal{O}(n_\Lambda^p N^{-1})$ and use (3.5). The evolution in time of the sublattice magnetisations is then found to be deterministic in the thermodynamic limit if

$$\lim_{N\to\infty} \frac{p}{\log N} = 0$$

Furthermore, condition (3.7) holds, since

$$\sum_{j=1}^{N} w_j(\boldsymbol{\sigma}) \Delta_{j\boldsymbol{\eta}}(\boldsymbol{\sigma}) = \tanh\left[\beta \sum_{\boldsymbol{\eta}'} p_{\boldsymbol{\eta}'} Q\left(\boldsymbol{\eta}; \boldsymbol{\eta}'\right) m_{\boldsymbol{\eta}'}\right] - m_{\boldsymbol{\eta}}$$

We may conclude that the evolution in time of the sublattice magnetisations is governed by the following autonomous set of differential equations:

$$\frac{d}{dt}m_{\boldsymbol{\eta}} = \tanh\left[\beta\sum_{\boldsymbol{\eta}'}p_{\boldsymbol{\eta}'}Q\left(\boldsymbol{\eta};\boldsymbol{\eta}'\right)m_{\boldsymbol{\eta}'}\right] - m_{\boldsymbol{\eta}} \tag{3.10}$$

The above procedure does not require symmetry of the interaction matrix. In the symmetric case $Q(\boldsymbol{x};\boldsymbol{y}) = Q(\boldsymbol{y};\boldsymbol{x})$ the system will approach equilibrium; if the kernel Q is positive definite this can be shown, for instance, by inspection of the Liapunov function $\mathcal{L}\{m_{\boldsymbol{\eta}}\}$:

$$\mathcal{L}\{m_{\boldsymbol{\eta}}\} = \frac{1}{2}\sum_{\boldsymbol{\eta}\boldsymbol{\eta}'}p_{\boldsymbol{\eta}}m_{\boldsymbol{\eta}}Q(\boldsymbol{\eta};\boldsymbol{\eta}')m_{\boldsymbol{\eta}'}p_{\boldsymbol{\eta}'} - \frac{1}{\beta}\sum_{\boldsymbol{\eta}}p_{\boldsymbol{\eta}}\log\cosh\left[\beta\sum_{\boldsymbol{\eta}'}Q(\boldsymbol{\eta};\boldsymbol{\eta}')m_{\boldsymbol{\eta}'}p_{\boldsymbol{\eta}'}\right]$$

which is bounded and obeys:

$$\frac{d}{dt}\mathcal{L} = -\sum_{\boldsymbol{\eta}\boldsymbol{\eta}'}\left[p_{\boldsymbol{\eta}}\frac{d}{dt}m_{\boldsymbol{\eta}}\right]Q(\boldsymbol{\eta};\boldsymbol{\eta}')\left[p_{\boldsymbol{\eta}'}\frac{d}{dt}m_{\boldsymbol{\eta}'}\right] \leq 0 \tag{3.11}$$

Note that from the sublattice magnetisations one can easily calculate the 'overlap' order parameters (2.6), which can be written as averages over the n_Λ^p sublattice magnetisations:

$$m_\mu(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{i=1}^{N}\xi_i^\mu\sigma_i = \sum_{\boldsymbol{\eta}}p_{\boldsymbol{\eta}}\eta_\mu m_{\boldsymbol{\eta}} \tag{3.12}$$

Whether or not, in turn, there exists an *autonomous* set of laws at the level of overlaps depends on the form of the kernel $Q(.;.)$.

Simple examples of relevant models of the type (3.8), the dynamics of which is for large N described by equation (3.10), are for instance the ones where one applies a non-linear operation Φ to the standard Hopfield interactions (2.3):

$$Q(\boldsymbol{x};\boldsymbol{y}) = \Phi(\boldsymbol{x}\cdot\boldsymbol{y}): \qquad J_{ij} \equiv \frac{1}{N}\Phi\left(\sum_{\mu\leq p}\xi_i^\mu\xi_j^\mu\right)$$

with $\Phi(0) = 0$ and $\Phi'(x) \geq 0$. This non-linearity could result from e.g. a clipping procedure,

$$\Phi(x) = \begin{cases} -K & \text{for} \quad x \leq K \\ x & \text{for} \quad -K < x < K \\ K & \text{for} \quad x \geq K \end{cases}$$

or from retaining only the *sign* of the original interactions (2.3):

$$\Phi(x) = \operatorname{sgn}(x)$$

It turns out that the effect of introducing such non-linearities is of a quantitative nature, giving rise only to little more than a rescaling of critical temperatures and storage capacities. We will not go into full details, but illustrate this statement with a simple example, by working out the $p = 2$ equations for randomly drawn pattern bits $\xi^\mu \in \{-1, 1\}$, where there are only four sublattices and where $p_{\boldsymbol{\eta}} = \frac{1}{4}$ for all $\boldsymbol{\eta}$. Using $\Phi(0) = 0$ and $\Phi(-x) = -\Phi(x)$ (as is the case for the above examples) we obtain:

$$\frac{d}{dt}m_{\boldsymbol{\eta}} = \tanh\left[\frac{1}{4}\beta\Phi(2)(m_{\boldsymbol{\eta}} - m_{-\boldsymbol{\eta}})\right] - m_{\boldsymbol{\eta}} \qquad (3.13)$$

which shows the choice made for the non-linearity $\Phi(x)$ only to show up as a rescaling of the temperature, at least for the simple case $p = 2$. From (3.13) we further obtain $\frac{d}{dt}(m_{\boldsymbol{\eta}}+m_{-\boldsymbol{\eta}}) = -(m_{\boldsymbol{\eta}}+m_{-\boldsymbol{\eta}})$. The system decays exponentially towards a state where, according to (3.12), all sublattices contribute equally to the overlaps: $m_{\boldsymbol{\eta}} = -m_{-\boldsymbol{\eta}}$ for all $\boldsymbol{\eta}$. If at $t = 0$ this is already the case, we simply find decoupled equations describing four infinite-range ferro-magnets.

Separable Models: Description at the Level of Overlaps. Equation (3.10) suggests that at the level of overlaps there will be, in turn, an autonomous set of dynamical laws if the kernel Q is bilinear[1], i.e. $Q(\boldsymbol{x};\boldsymbol{y}) = \sum_{\mu\nu} x_\mu A_{\mu\nu} y_\nu$, or:

$$J_{ij} \equiv \frac{1}{N}\sum_{\mu\nu=1}^{p} \xi_i^\mu A_{\mu\nu} \xi_j^\nu \qquad\qquad \boldsymbol{\xi}_i \equiv (\xi_i^1, \ldots, \xi_i^p) \qquad (3.14)$$

Now the components ξ_i^μ need not be drawn from a finite discrete set, as we will see (as long as they do not depend on N). The Hopfield model corresponds to choosing $A_{\mu\nu} \equiv \delta_{\mu\nu}$ and $\xi_i^\mu \in \{-1, 1\}$. The alignment fields h_k can now be written in terms of the overlap order parameters m_μ:

$$m_\mu(\boldsymbol{\sigma}) \equiv \frac{1}{N}\sum_{i=1}^{N} \xi_i^\mu \sigma_i \qquad h_k(\boldsymbol{\sigma}) = \boldsymbol{\xi}_k \cdot A\boldsymbol{m}(\boldsymbol{\sigma}) \qquad \boldsymbol{m} \equiv (m_1, \ldots, m_p) \qquad (3.15)$$

Since for the present choice of macroscopic variables we find $\Delta_{j\mu} = \mathcal{O}(N^{-1})$, the evolution in time of the overlap vector $\boldsymbol{m}$ becomes deterministic in the thermodynamic limit if (according to (3.5)):

$$\lim_{N\to\infty} \frac{p}{\sqrt{N}} = 0$$

Again condition (3.7) holds, since

$$\sum_{j=1}^{N} w_j(\boldsymbol{\sigma})\Delta_{j\mu}(\boldsymbol{\sigma}) = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{\xi}_k \tanh\left[\beta\boldsymbol{\xi}_k \cdot A\boldsymbol{m}\right] - \boldsymbol{m}$$

In the thermodynamic limit the evolution in time of the overlap vector $\boldsymbol{m}$ is governed by an autonomous set of differential equations; if the vectors $\boldsymbol{\xi}_k$ are drawn at random according to some distribution $\rho(\boldsymbol{\xi})$ these dynamical laws become:

$$\frac{d}{dt}\boldsymbol{m} = \langle \boldsymbol{\xi} \tanh\left[\beta\boldsymbol{\xi} \cdot A\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}} - \boldsymbol{m} \qquad \langle \Phi(\boldsymbol{\xi})\rangle_{\boldsymbol{\xi}} \equiv \int d\boldsymbol{\xi}\ \rho(\boldsymbol{\xi})\Phi(\boldsymbol{\xi}) \qquad (3.16)$$

[1] Strictly speaking, it is already sufficient to have a kernel which is linear in $\boldsymbol{y}$ only, i.e. $Q(\boldsymbol{x};\boldsymbol{y}) = \sum_\nu f_\nu(\boldsymbol{x}) y_\nu$

Again symmetry of the interaction matrix is not required. For specific non-symmetric choices for the matrix A stable limit-cycle solutions of (3.16) can be found. In the symmetric case $A_{\mu\nu} = A_{\nu\mu}$ the system will approach equilibrium; the Liapunov function (3.11) for positive definite matrices A now becomes:

$$\mathcal{L}\{m_\mu\} = \frac{1}{2}\boldsymbol{m}\cdot A\boldsymbol{m} - \frac{1}{\beta}\langle \log\cosh\left[\beta\boldsymbol{\xi}\cdot A\boldsymbol{m}\right]\rangle_{\boldsymbol{\xi}}$$

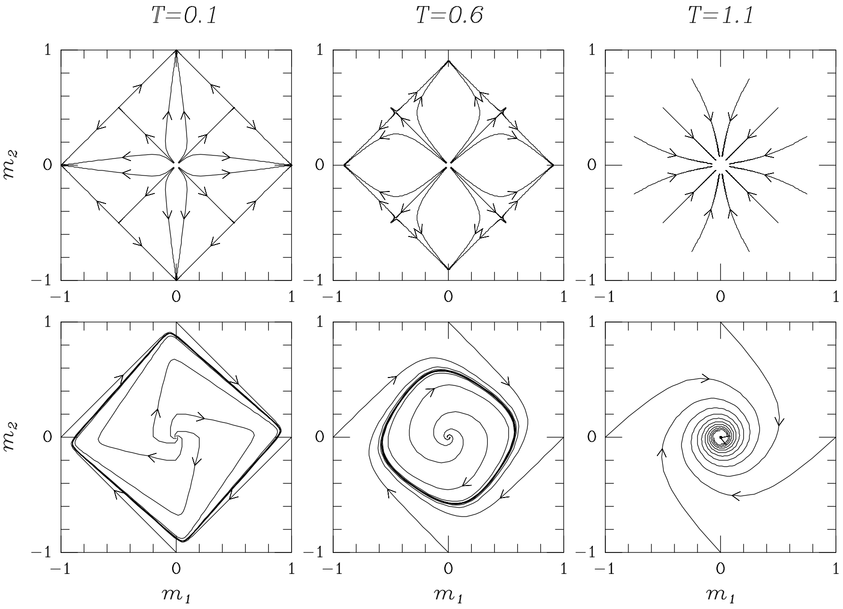


Figure 3.1: Flow diagrams obtained by numerically solving the deterministic overlap equations for $p = 2$. Upper row: $A_{\mu\nu} = \delta_{\mu\nu}$ (the Hopfield model); lower row: $A = \left(\begin{smallmatrix}1 & 1\\ -1 & 1\end{smallmatrix}\right)$ (for both models the critical temperature is $T_c = 1$).

Figure 3.1 shows in the m_1, m_2-plane the result of solving the macroscopic laws (3.16) numerically for $p = 2$, randomly drawn pattern bits $\xi_i^\mu \in \{-1, 1\}$, and two choices of the matrix A. The first choice (upper row) corresponds to the Hopfield model; as the temperature increases the amplitudes of the four attractors (corresponding to the two patterns $\boldsymbol{\xi}^\mu$ and their mirror images $-\boldsymbol{\xi}^\mu$) continuously decrease, until at the critical temperature $T_c = 1$ they merge into the trivial attractor $\boldsymbol{m} = (0, 0)$. The second choice corresponds to a non-symmetric model (i.e. without detailed balance); at the macroscopic level of desciption (at finite timescales) the system clearly does not approach equilibrium; macroscopic order now manifests itself in the form of a limit-cycle (provided the temperature T is below the critical temperature $T_c = 1$ where this limit-cycle is destroyed). To what extent the laws (3.16) are in agreement with the result of performing the actual simulations in finite systems is illustrated in figure 3.2.

As a second simple application of the flow equations (3.16) we turn to the relaxation times corresponding to the attractors of the Hopfield model (where $A_{\mu\nu} = \delta_{\mu\nu}$). Expanding (3.16)

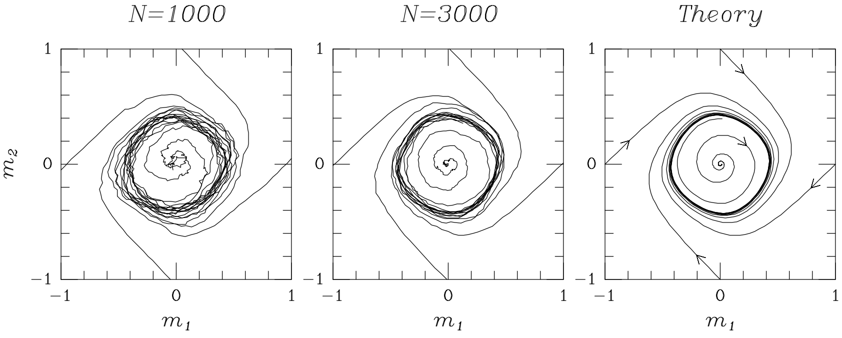


Figure 3.2: Comparison between simulation results for finite systems ($N = 1000$ and $N = 3000$) and the analytical prediction (flow equations) with respect to the evolution of the order parameters (m_1, m_2); $p = 2$, $T = 0.8$ and $A = \left(\begin{smallmatrix} 1 & 1 \\ -1 & 1 \end{smallmatrix}\right)$.

near a stable fixed-point $\boldsymbol{m}^*$, i.e. $\boldsymbol{m} = \boldsymbol{m}^* + \boldsymbol{x}$ with $|\boldsymbol{x}| \ll 1$, gives the linearised equation

$$\frac{d}{dt}x_\mu = \left[\beta \sum_\nu \langle \xi_\mu \xi_\nu \tanh[\beta \boldsymbol{\xi} \cdot \boldsymbol{m}^*] \rangle_{\boldsymbol{\xi}} - \delta_{\mu\nu}\right] x_\nu + \mathcal{O}(\boldsymbol{x}^2) \tag{3.17}$$

The Jacobian of (3.16), which determines the linearised equation (3.17), turns out to be *minus* the curvature matrix (2.15) of the free energy surface at the fixed-point. The asymptotic relaxation towards any stable attractor is therefore exponential, with a characteristic time τ given by the inverse of the smallest eigenvalue of the curvature matrix (2.15). If, in particular, for the fixed point $\boldsymbol{m}^*$ we substitute an n-mixture state (2.12), and transform (3.17) to the basis where the corresponding curvature matrix $D^{(n)}$ (2.16) (with eigenvalues D^n_λ) is diagonal, $\boldsymbol{x} \to \tilde{\boldsymbol{x}}$, we obtain

$$\tilde{x}_\lambda(t) = \tilde{x}_\lambda(0) e^{-tD^n_\lambda} + \ldots$$

so $\tau^{-1} = \min_\lambda D^n_\lambda$, which we have already calculated in determining the character of the saddle-points of the free-energy surface. The result is shown in figure 3.3. The relaxation time for the n-mixture attractors decreases monotonically with the degree of mixing n, for any temperature.

At the transition where a macroscopic state $\boldsymbol{m}^*$ ceases to correspond to a local minimum of the free energy surface, it also destabilises in terms of the linearised dynamic equation (3.17) (as it should). The Jacobian develops a zero eigenvalue, the relaxation time diverges, and the long-time behaviour is no longer obtained from the linearised equation. This gives rise to critical slowing down (power law relaxation as opposed to exponential relaxation). For instance, at the transition temperature $T_c = 1$ for the $n = 1$ (pure) state, we find by expanding (3.16):

$$\frac{d}{dt}m_\mu = m_\mu \left[\frac{2}{3}m_\mu^2 - \boldsymbol{m}^2\right] + \mathcal{O}(\boldsymbol{m}^5)$$

which gives rise to a relaxation towards the trivial fixed-point of the form $\boldsymbol{m} \sim t^{-\frac{1}{2}}$.

If one is willing to pay the price of restricting oneself to the limited class of models (3.14) (as opposed to the more general class (3.8)) and to the more global level of description in terms

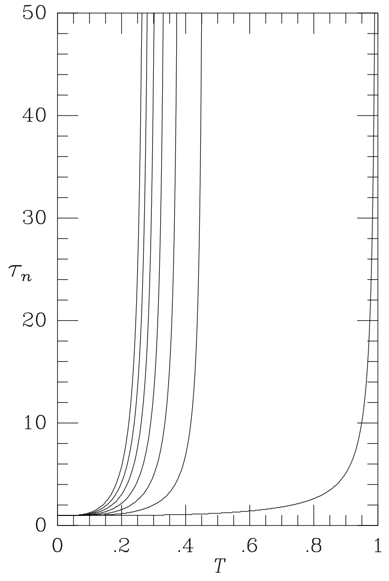


Figure 3.3: Asymptotic relaxation times τ_n of the mixture states of the Hopfield model as a function of temperature. From bottom to top: $n = 1, 3, 5, 7, 9, 11, 13$.

of p overlap parameters m_μ instead of n_Λ^p sublattice magnetisations $m_{\boldsymbol{\eta}}$, then there are two rewards. Firstly there will be no restrictions on the stored quantities ξ_i^μ (for instance, they are allowed to be real-valued); secondly the number p of patterns stored can be much larger for the deterministic autonomous dynamical laws to hold ($p << \sqrt{N}$ instead of $p << \log N$, which from a biological point of view is too restrictive.

3.2 Parallel Dynamics

We now turn to the case of parallel dynamics, i.e. the discrete-time stochastic microscopic laws (1.3). The evolution of the macroscopic probability distribution will now be described by discrete mappings, in stead of differential equations.

The Toy Model. Let us first see what happens to our toy model (3.1). As before we try to describe the dynamics at the (macroscopic) level of the quantity $m(\boldsymbol{\sigma}) = \frac{1}{N}\sum_k \xi_k \sigma_k$. The evolution of the macroscopic probability distribution $\mathcal{P}_t[m]$ is obtained by inserting (1.3):

$$\mathcal{P}_{t+1}[m] = \sum_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} \delta\left[m - m(\boldsymbol{\sigma})\right] W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_t(\boldsymbol{\sigma}')$$

$$= \int dm'\ \tilde{W}_t\left[m, m'\right] \mathcal{P}_t[m'] \tag{3.18}$$

with

$$\tilde{W}_t\left[m, m'\right] = \frac{\sum_{\boldsymbol{\sigma}\boldsymbol{\sigma}'} \delta\left[m - m(\boldsymbol{\sigma})\right] \delta\left[m' - m(\boldsymbol{\sigma}')\right] W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right] p_t(\boldsymbol{\sigma}')}{\sum_{\boldsymbol{\sigma}'} \delta\left[m' - m(\boldsymbol{\sigma}')\right] p_t(\boldsymbol{\sigma}')}$$

We now insert the expression (1.4) for the transition probabilities and the local fields. Due to the fact that the fields depend on the microscopic state $\boldsymbol{\sigma}$ only through $m(\boldsymbol{\sigma})$, the microscopic distribution $p_t(\boldsymbol{\sigma})$ drops out of the above expression for the kernel $\tilde{W}_t$ which thereby loses its explicit time-dependence, $\tilde{W}_t\left[m,m'\right] \to \tilde{W}\left[m,m'\right]$:

$$\tilde{W}\left[m,m'\right] = e^{-\sum_i \log\cosh(\beta J m'\eta_i)} \langle \delta\left[m - m(\boldsymbol{\sigma})\right] e^{\beta J m' \sum_i \eta_i\sigma_i} \rangle_{\boldsymbol{\sigma}}$$

(with $\langle\ldots\rangle_{\boldsymbol{\sigma}} = 2^{-N}\sum_{\boldsymbol{\sigma}}\ldots$). Inserting the integral representation for the δ-function allows us to perform the spin-average:

$$\tilde{W}\left[m,m'\right] = \left[\frac{\beta N}{2\pi}\right] \int dk\ e^{N\Psi(m,m',k)}$$

$$\Psi = i\beta k m + \langle \log\cosh\beta[J\eta m' - ik\xi]\rangle_{\eta,\xi} - \langle\log\cosh\beta[J\eta m']\rangle_\eta$$

Since $\tilde{W}\left[m,m'\right]$ is (by construction) normalised, $\int dm\ \tilde{W}\left[m,m'\right] = 1$, we find that for $N \to \infty$ the expectation value with respect to $\tilde{W}\left[m,m'\right]$ of any sufficiently smooth function $f(m)$ will be determined only by the value $m^*(m')$ of m in the relevant saddle-point of Ψ:

$$\int dm\ f(m)\tilde{W}\left[m,m'\right] = \frac{\int dm dk\ f(m) e^{N\Psi(m,m',k)}}{\int dm dk\ e^{N\Psi(m,m',k)}} \to f(m^*(m')) \qquad (N\to\infty)$$

Variation of Ψ with respect to k and m gives the two saddle-point equations:

$$m = \langle \xi\tanh\beta[J\eta m' - \xi k]\rangle_{\eta,\xi} \qquad k = 0$$

We may now conclude that $\lim_{N\to\infty}\tilde{W}\left[m,m'\right] = \delta\left[m - m^*(m')\right]$ with $m^*(m') = \langle\xi\tanh(\beta J\eta m')\rangle_{\eta,\xi}$, and that the macroscopic equation (3.18) becomes:

$$\mathcal{P}_{t+1}[m] = \int dm'\ \delta\left[m - \langle\xi\tanh(\beta J\eta m')\rangle_{\eta\xi}\right]\mathcal{P}_t[m'] \qquad (N\to\infty)$$

This relation, of course, describes deterministic evolution. If at $t = 0$ we know m exactly, this will remain the case for finite timescales and m will evolve according to a discrete version of the sequential dynamics flow equation (3.2):

$$m_{t+1} = \langle\xi\tanh[\beta J\eta m_t]\rangle_{\eta,\xi} \tag{3.19}$$

Arbitrary Synaptic Interactions. We now try to generalise the above approach to less trivial classes of models. As for the sequential case we will find in the limit $N \to \infty$ closed deterministic evolution equations for a more general set of intensive macroscopic state variables $\boldsymbol{\Omega}(\boldsymbol{\sigma}) = (\Omega_1(\boldsymbol{\sigma}),\ldots,\Omega_n(\boldsymbol{\sigma})$ if the local alignment fields (1.2) depend on the microscopic state $\boldsymbol{\sigma}$ only through the values of $\boldsymbol{\Omega}(\boldsymbol{\sigma})$ in the limit of infinitely large networks and if the number n of these state variables necessary to do so is not too large.

The evolution of the ensemble probability of finding the system in macroscopic state $\boldsymbol{\Omega}$,

$$\mathcal{P}_t\left[\boldsymbol{\Omega}\right] \equiv \sum_{\boldsymbol{\sigma}} p_t(\boldsymbol{\sigma})\delta\left[\boldsymbol{\Omega} - \boldsymbol{\Omega}(\boldsymbol{\sigma})\right]$$

is obtained by inserting the Markov equation (1.3) with the transition probabilities (1.4) and the local fields (1.2):

$$\mathcal{P}_{t+1}\left[\mathbf{\Omega}\right] = \int d\mathbf{\Omega}'\ \tilde{W}_t\left[\mathbf{\Omega},\mathbf{\Omega}'\right]\mathcal{P}_t\left[\mathbf{\Omega}'\right] \tag{3.20}$$

$$\begin{aligned}\tilde{W}_t\left[\mathbf{\Omega},\mathbf{\Omega}'\right] &= \frac{\sum_{\boldsymbol{\sigma}\boldsymbol{\sigma}'}\delta\left[\mathbf{\Omega}-\mathbf{\Omega}(\boldsymbol{\sigma})\right]\delta\left[\mathbf{\Omega}'-\mathbf{\Omega}(\boldsymbol{\sigma}')\right]W\left[\boldsymbol{\sigma};\boldsymbol{\sigma}'\right]p_t(\boldsymbol{\sigma}')}{\sum_{\boldsymbol{\sigma}'}\delta\left[\mathbf{\Omega}'-\mathbf{\Omega}(\boldsymbol{\sigma}')\right]p_t(\boldsymbol{\sigma}')} \\ &= \langle\delta\left[\mathbf{\Omega}-\mathbf{\Omega}(\boldsymbol{\sigma})\right]\langle e^{\sum_i[\beta\sigma_i h_i(\boldsymbol{\sigma}')-\log\cosh(\beta h_i(\boldsymbol{\sigma}'))]}\rangle_{\mathbf{\Omega}';t}\rangle_{\boldsymbol{\sigma}}\end{aligned} \tag{3.21}$$

with an ordinary homogeneous spin-average $\langle\ldots\rangle_{\boldsymbol{\sigma}} = 2^{-N}\sum_{\boldsymbol{\sigma}}\ldots$, and a sub-shell (or conditional) spin-average, defined as

$$\langle f(\boldsymbol{\sigma})\rangle_{\mathbf{\Omega};t} = \frac{\sum_{\boldsymbol{\sigma}} f(\boldsymbol{\sigma})\delta\left[\mathbf{\Omega}-\mathbf{\Omega}(\boldsymbol{\sigma})\right]p_t(\boldsymbol{\sigma})}{\sum_{\boldsymbol{\sigma}}\delta\left[\mathbf{\Omega}-\mathbf{\Omega}(\boldsymbol{\sigma})\right]p_t(\boldsymbol{\sigma})}$$

It is clear from (3.21) that in order to find autonomous macroscopic equations, i.e. for the microscopic distribution $p_t(\boldsymbol{\sigma})$ to drop out, the local alignment fields must depend on the microscopic state $\boldsymbol{\sigma}$ only through the macroscopic quantities $\mathbf{\Omega}(\boldsymbol{\sigma})$:

$$h_i(\boldsymbol{\sigma}) = h_i[\mathbf{\Omega}(\boldsymbol{\sigma})]$$

In this case $\tilde{W}_t$ loses its explicit time-dependence, $\tilde{W}_t\left[\mathbf{\Omega},\mathbf{\Omega}'\right] \to \tilde{W}\left[\mathbf{\Omega},\mathbf{\Omega}'\right]$. Inserting integral representations for the δ-functions leads to:

$$\tilde{W}\left[\mathbf{\Omega},\mathbf{\Omega}'\right] = \left[\frac{\beta N}{2\pi}\right]^n \int d\boldsymbol{K}\ e^{N\Psi(\mathbf{\Omega},\mathbf{\Omega}',\boldsymbol{K})}$$

$$\Psi = i\beta\boldsymbol{K}\cdot\mathbf{\Omega} + \frac{1}{N}\log\langle e^{\beta\left[\sum_i \sigma_i h_i[\mathbf{\Omega}'] - iN\boldsymbol{K}\cdot\mathbf{\Omega}(\boldsymbol{\sigma})\right]}\rangle_{\boldsymbol{\sigma}} - \frac{1}{N}\sum_i \log\cosh[\beta h_i[\mathbf{\Omega}']$$

Using the normalisation $\int d\mathbf{\Omega}\ \tilde{W}\left[\mathbf{\Omega},\mathbf{\Omega}'\right] = 1$, we can write expectation values with respect to $\tilde{W}\left[\mathbf{\Omega},\mathbf{\Omega}'\right]$ of macroscopic quantities $f[\mathbf{\Omega}]$ as

$$\int d\mathbf{\Omega}\ f[\mathbf{\Omega}]\tilde{W}\left[\mathbf{\Omega},\mathbf{\Omega}'\right] = \frac{\int d\mathbf{\Omega} d\boldsymbol{K}\ f[\mathbf{\Omega}]e^{N\Psi(\mathbf{\Omega},\mathbf{\Omega}',\boldsymbol{K})}}{\int d\mathbf{\Omega} d\boldsymbol{K}\ e^{N\Psi(\mathbf{\Omega},\mathbf{\Omega}',\boldsymbol{K})}} \tag{3.22}$$

For steepest descent to apply in determining the leading order in N of the average (3.22), at this stage we encounter restrictions on the number n of our macroscopic quantities, since n determines the dimension of the integrations concerned. The restrictions can be found by expanding Ψ around its maximum Ψ^*. After defining $\boldsymbol{x} = (\mathbf{\Omega},\boldsymbol{K})$, of dimension $2n$, and after translating the location of the maximum to the origin, one has

$$\Psi(\boldsymbol{x}) = \Psi^* - \frac{1}{2}\sum_{\mu\nu} x_\mu x_\nu K_{\mu\nu} + \sum_{\mu\nu\rho} x_\mu x_\nu x_\rho L_{\mu\nu\rho} + \mathcal{O}(\boldsymbol{x}^4)$$

so

$$\int d\boldsymbol{x}\ e^{N\Psi(\boldsymbol{x})} = e^{N\Psi^*}\int d\boldsymbol{x}\ e^{-\frac{1}{2}N\boldsymbol{x}\cdot K\boldsymbol{x} + N\sum_{\mu\nu\rho} x_\mu x_\nu x_\rho L_{\mu\nu\rho} + \mathcal{O}(N\boldsymbol{x}^4)}$$

$$= e^{N\Psi^*} N^{-n} \int d\boldsymbol{y}\; e^{-\frac{1}{2}\boldsymbol{y}\cdot K\boldsymbol{y}} \left[1+\mathcal{O}\left(\frac{n^2}{N}\right)\right]$$

$$\frac{1}{N}\log\int d\boldsymbol{x}\; e^{N\Psi(\boldsymbol{x})} = \Psi^* + \frac{n}{N}\log\left[\frac{2\pi}{N}\right] - \frac{1}{2N}\sum_\mu \log K_\mu + \mathcal{O}\left(\frac{n}{N}\right)^2$$

where $\{K_\mu\}$ are the (positive) eigenvalues of the curvature matrix K at the minimum of Ψ. Since, by definition, these eigenvalues scale with the dimension n as $K_\mu \sim n$ at the most, we find the condition

$$\lim_{N\to\infty} n\log N/N = 0: \qquad \lim_{N\to\infty}\int d\boldsymbol{\Omega}\; f[\boldsymbol{\Omega}]\tilde{W}\left[\boldsymbol{\Omega},\boldsymbol{\Omega}'\right] = f\left[\boldsymbol{\Omega}^*(\boldsymbol{\Omega}')\right]$$

where $\boldsymbol{\Omega}^*(\boldsymbol{\Omega}')$ denotes the value of $\boldsymbol{\Omega}$ in the saddle-point where Ψ is minimised.

Variation of Ψ with respect to $\boldsymbol{\Omega}$ and $\boldsymbol{K}$ gives the saddle-point equations:

$$\boldsymbol{\Omega} = \frac{\langle \boldsymbol{\Omega}(\boldsymbol{\sigma}) e^{\beta\left[\sum_i \sigma_i h_i[\boldsymbol{\Omega}'] - iN\boldsymbol{K}\cdot\boldsymbol{\Omega}(\boldsymbol{\sigma})\right]}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\left[\sum_i \sigma_i h_i[\boldsymbol{\Omega}'] - iN\boldsymbol{K}\cdot\boldsymbol{\Omega}(\boldsymbol{\sigma})\right]}\rangle_{\boldsymbol{\sigma}}} \qquad \boldsymbol{K} = 0$$

We may now conclude that $\lim_{N\to\infty}\tilde{W}\left[\boldsymbol{\Omega},\boldsymbol{\Omega}'\right] = \delta\left[\boldsymbol{\Omega} - \boldsymbol{\Omega}^*(\boldsymbol{\Omega}')\right]$, with

$$\boldsymbol{\Omega}^*(\boldsymbol{\Omega}') = \frac{\langle \boldsymbol{\Omega}(\boldsymbol{\sigma}) e^{\beta\sum_i \sigma_i h_i[\boldsymbol{\Omega}']}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\sum_i \sigma_i h_i[\boldsymbol{\Omega}']}\rangle_{\boldsymbol{\sigma}}}$$

and that for $N \to \infty$ the macroscopic equation (3.20) becomes:

$$\mathcal{P}_{t+1}\left[\boldsymbol{\Omega}\right] = \int d\boldsymbol{\Omega}'\; \delta\left[\boldsymbol{\Omega} - \frac{\langle \boldsymbol{\Omega}(\boldsymbol{\sigma}) e^{\beta\sum_i \sigma_i h_i[\boldsymbol{\Omega}']}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\sum_i \sigma_i h_i[\boldsymbol{\Omega}']}\rangle_{\boldsymbol{\sigma}}}\right] \mathcal{P}_t\left[\boldsymbol{\Omega}'\right]$$

This relation again describes deterministic evolution. If at $t = 0$ we know $\boldsymbol{\Omega}$ exactly, this will remain the case for finite timescales and $\boldsymbol{\Omega}$ will evolve according to

$$\boldsymbol{\Omega}(t+1) = \frac{\langle \boldsymbol{\Omega}(\boldsymbol{\sigma}) e^{\beta\sum_i \sigma_i h_i[\boldsymbol{\Omega}(t)]}\rangle_{\boldsymbol{\sigma}}}{\langle e^{\beta\sum_i \sigma_i h_i[\boldsymbol{\Omega}(t)]}\rangle_{\boldsymbol{\sigma}}} \tag{3.23}$$

As with the sequential case, in taking the limit $N \to \infty$ we have to keep in mind that the resulting deterministic theory applies to finite t, and that for sufficiently large times terms of higher order in N do come into play. If compared to the sequential case the restriction $n\log n/N \to 0$ on the number of macroscopic state variables is less severe, this indicates that in the sequential case we can probably further sharpen our statements, should the need arise.

Finally, for macroscopic quantities $\boldsymbol{\Omega}(\boldsymbol{\sigma})$ which are linear in $\boldsymbol{\sigma}$, the remaining spin-average becomes trivial, so that:

$$\boldsymbol{\Omega}(\boldsymbol{\sigma}) = \frac{1}{N}\sum_i \boldsymbol{\omega}_i\sigma_i: \qquad \boldsymbol{\Omega}(t+1) = \lim_{N\to\infty}\frac{1}{N}\sum_i \boldsymbol{\omega}_i \tanh\left[\beta h_i[\boldsymbol{\Omega}(t)]\right] \tag{3.24}$$

Separable models: Sublattice Magnetisations and Overlaps. The separable class of attractor models (3.8), described at the level of sublattice magnetisations (3.9), indeed has the desired property that all local fields can be written in terms of the macroscopic state variables (the sublattice magnetisations) only. What remains is the restriction on the number n of these macroscopic state variables, to ensure deterministic evolution. If all relative sublattice sizes $p_{\boldsymbol{\eta}}$ are of the same order in N (as for randomly drawn patterns) this condition again translates into

$$\lim_{N\to\infty}\frac{p}{\log N}=0$$

Since the sublattice magnetisations are linear functions of the spins, their evolution in time is now governed by equation (3.24), which acquires the form:

$$m_{\boldsymbol{\eta}}(t+1)=\tanh\left[\beta\sum_{\boldsymbol{\eta}'}p_{\boldsymbol{\eta}'}Q\left(\boldsymbol{\eta};\boldsymbol{\eta}'\right)m_{\boldsymbol{\eta}'}(t)\right] \tag{3.25}$$

As for sequential dynamics, symmetry of the interaction matrix does not play a role.

At the more global level of overlaps $\boldsymbol{m}(\boldsymbol{\sigma})$ (2.6) we obtain autonomous deterministic mappings if the local fields (1.2) can be expressed in terms if $\boldsymbol{m}(\boldsymbol{\sigma})$ only, as for the models (3.14) (or, more generally, for all models in which the interactions are of the form $J_{ij}=\sum_{\mu\leq p}f_{i\mu}\xi_j^\mu$), and with a restriction on the number p of embedded patterns:

$$\lim_{N\to\infty}\frac{p\log p}{N}=0$$

For the class of bi-linear models (3.14), the evolution in time of the overlap vector $\boldsymbol{m}$ (which depends linearly on the spin variables) is governed by (3.24), which now translates into the iterative map:

$$\boldsymbol{m}(t+1)=\langle\boldsymbol{\xi}\tanh\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{A}\boldsymbol{m}(t)\right]\rangle_{\boldsymbol{\xi}}\qquad\langle\Phi(\boldsymbol{\xi})\rangle_{\boldsymbol{\xi}}\equiv\int d\boldsymbol{\xi}\;\rho(\boldsymbol{\xi})\Phi(\boldsymbol{\xi}) \tag{3.26}$$

Again symmetry of the interaction matrix is not required. For parallel dynamics it is far more difficult than for sequential dynamics to construct Liapunov functions and prove that the macroscopic equations for symmetric systems evolve towards a stable fixed-point (as one would expect), but it can still be done. For non-symmetric systems the final macroscopic equations can in principle display all the interesting, but complicated, phenomena of non-conservative non-linear systems. Nevertheless, it is also not uncommon that the macroscopic equations for non-symmetric systems can be mapped by a time-dependent transformation onto the equations for related symmetric systems (mostly variants of the original Hopfield model), such that a thorough analysis is possible. An example of such a model is given below.

As an example we show in figure 3.4 as functions of time the values of the overlap order parameters $\{m_\mu\}$ for $p=10$ and $T=0.5$, resulting from numerical iteration of the macroscopic laws (3.26) for the model

$$J_{ij}=\frac{\nu}{N}\sum_\mu\xi_i^\mu\xi_j^\mu+\frac{1-\nu}{N}\sum_\mu\xi_i^{\mu+1}\xi_j^\mu\qquad(\mu:\ \bmod p)$$

or:

$$A_{\lambda\rho}=\nu\delta_{\lambda\rho}+(1-\nu)\delta_{\lambda,\rho+1}\qquad(\lambda,\rho:\ \bmod p)$$

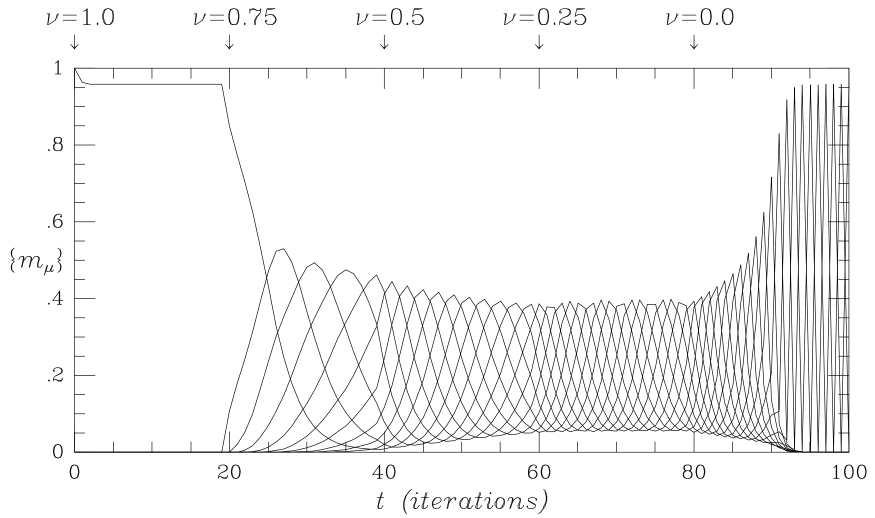


Figure 3.4: Evolution of overlaps $m_\mu(\boldsymbol{\sigma})$, obtained by numerical iteration of the parallel dynamics deterministic macroscopic map, for the bi-linear attractor model $J_{ij} = \frac{\nu}{N}\sum_{\mu\nu}\xi_i^\mu\xi_j^\nu + \frac{1-\nu}{N}\sum_{\mu\nu}\xi_i^{\mu+1}\xi_j^\nu$, with $p = 10$ and $T = 0.5$.

with randomly drawn pattern bits $\xi_i^\mu \in \{-1,1\}$ The initial state is choosen to be the pure state $m_\mu = \delta_{\mu,1}$. At intervals of $\Delta t = 20$ iterations the parameter ν is reduced in $\Delta\nu = 0.25$ steps from $\nu = 1$ (where one recovers the symmetric Hopfield model) to $\nu = 0$ (where one obtains a non-symmetric model which processes the p embedded patterns in strict sequential order as a period-p limit-cycle).

The analysis of the equations (3.26) for the pure sequence processing case $\nu = 0$ is greatly simplified by mapping the model onto the ordinary ($\nu = 1$) Hopfield model, using the index permutation symmetries of the pattern distribution, in the following way[2] (all pattern indices are periodic, mod p):

$$\begin{aligned} m_\mu(t) = M_{\mu-t}(t):\quad M_\mu(t+1) &= \langle \xi_{\mu+t+1}\tanh\left[\beta\textstyle\sum_\rho \xi_{\rho+1}M_{\rho-t}(t)\right]\rangle_{\boldsymbol{\xi}} \\ &= \langle \xi_{\mu+t+1}\tanh\left[\beta\textstyle\sum_\rho \xi_{\rho+t+1}M_\rho(t)\right]\rangle_{\boldsymbol{\xi}} \\ &= \langle \xi_\mu \tanh\left[\beta\boldsymbol{\xi}\cdot\boldsymbol{M}(t)\right]\rangle_{\boldsymbol{\xi}} \end{aligned}$$

From this mapping we can immediately infer, in particular, that to each stable macroscopic fixed-point attractor of the original Hopfield model corresponds a stable period-p macroscopic limit-cycle attractor in the $\nu = 1$ sequence processing model (e.g. pure states ↔ pure sequences, mixture states ↔ mixture sequences), with identical amplitude as a function of temperature. Figure 3.4 shows for $\nu = 0$ (i.e. $t > 80$) a relaxation towards such a pure sequence.

[2] The mapping discussed here is a special case of a more general duality relating all trajectories $\{\boldsymbol{m}(t)\}$ obtained by iterating the macroscopic laws (3.26) for a given value of the parameter ν to all trajectories $\{\boldsymbol{m}(t)\}$ obtained for the parameter value $1-\nu$

Chapter 4

The Dynamics of Learning in Perceptrons

4.1 Introduction

In this chapter we study the dynamics of supervised learning in artificial neural networks. The basic scenario is as follows. A 'student' neural network executes a certain known operation $S : D \to R$, which is parametrised by a vector $\boldsymbol{J}$, usually representing synaptic weights and/or neuronal thresholds. Here D denotes the set of all possible 'questions' and R denotes the set of all possible 'answers'. The student is being trained to emulate a given 'teacher', which executes some as yet unknown operation $T : D \to R$. In order to achieve the objective the student network S tries gradually to improve its performance by adapting its parameters $\boldsymbol{J}$ according to an iterative procedure, using only examples of input vectors (or 'questions') $\boldsymbol{\xi} \in \Re^N$ which are drawn at random from a fixed training set $\tilde{D} \subseteq D$ of size $|\tilde{D}|$, and the corresponding values of the teacher outputs $T(\boldsymbol{\xi})$ (the 'correct answers'). The iterative procedure (the 'learning rule') is not allowed to involve any further knowledge of the operation T. As far as the student is concerned the teacher is an 'oracle', or 'black box'; the only information available about the inner workings of the black box is contained in the various answers $T(\boldsymbol{\xi})$ it provides. See figure 4.1. For simplicity we will assume each 'question' $\boldsymbol{\xi}$ to be equally likely to occur (generalization of what follows to the case where the questions $\boldsymbol{\xi}$ carry non-uniform probabilities or probability densities $p(\boldsymbol{\xi})$ is straightforward).

We will consider the following class of learning rules, i.e. of recipes for the iterative modification of the student's control parameters $\boldsymbol{J}$, which we will refer to as on-line learning rules. An input vector $\boldsymbol{\xi}(t)$ is drawn independently at each iteration step from the training set $\tilde{D}$, followed by a modification of the control parameters $\boldsymbol{J}$:

$$\boldsymbol{J}(t+1) = \boldsymbol{J}(t) + \boldsymbol{F}\left[\boldsymbol{\xi}(t), \boldsymbol{J}(t), T(\boldsymbol{\xi}(t))\right] \tag{4.1}$$

The integer variable $t = 0, 1, 2, 3, \ldots$ labels the iteration steps. Since this process is stochastic (Markovian) we introduce the probability density $\hat{p}_t(\boldsymbol{J})$ to find parameter vector $\boldsymbol{J}$ at discrete iteration step t. In terms of this microscopic probability density the processes (4.1) can be written as:

$$\hat{p}_{t+1}(\boldsymbol{J}) = \int d\boldsymbol{J}' \; W[\boldsymbol{J}; \boldsymbol{J}']\hat{p}_t(\boldsymbol{J}') \tag{4.2}$$

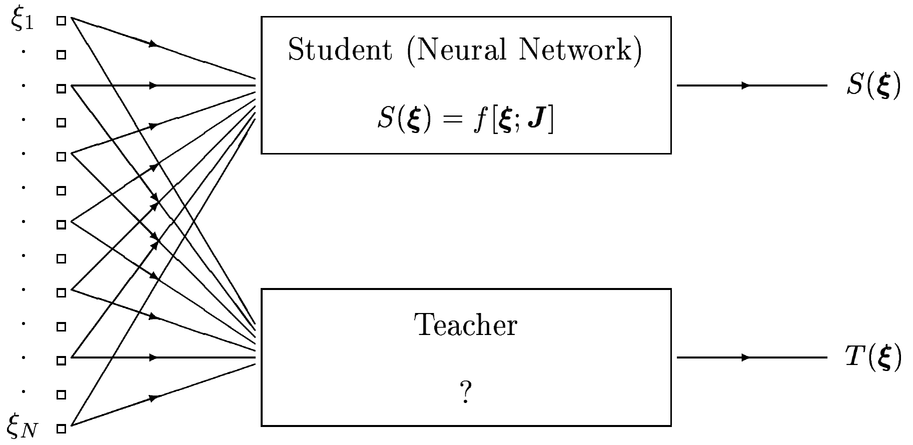


Figure 4.1: The general scenario of supervised learning: a 'student network' S is being 'trained' to perform an operation $T : D \rightarrow R$ by updating its control parameters $\boldsymbol{J}$ according to an iterative procedure, the 'learning rule'. This rule is allowed to make use only of examples of 'question/answer pairs' $(\boldsymbol{\xi}, T(\boldsymbol{\xi}))$, where $\boldsymbol{\xi} \in \tilde{D} \subseteq D$. The actual 'teacher operation' T that generated the answers $T(\boldsymbol{\xi})$, on the other hand, cannot be observed directly. The goal is to arrive at a situation where $S(\boldsymbol{\xi}) = T(\boldsymbol{\xi})$ for all $\boldsymbol{\xi} \in D$.

with the transition probability density

$$W[\boldsymbol{J}; \boldsymbol{J}'] = \langle \delta \left\{ \boldsymbol{J} - \boldsymbol{J}' - \boldsymbol{F} \left[\boldsymbol{\xi}, \boldsymbol{J}', T(\boldsymbol{\xi}) \right] \right\} \rangle_{\tilde{D}} \tag{4.3}$$

(in which $\delta[z]$ denotes the delta-distribution). The advantage of using on-line learning rules is a reduction in the amount of calculations that have to be done at each iteration step; the price paid for this reduction is the presence of fluctuations, with as yet unknown impact on the performance of the system.

We will denote averages over the probability density $\hat{p}_t(\boldsymbol{J})$, averages over the full set D of possible input vectors and averages over the training set $\tilde{D}$ in the following way:

$$\langle g(\boldsymbol{J}) \rangle = \int d\boldsymbol{J}\ \hat{p}_t(\boldsymbol{J}) g(\boldsymbol{J}) \qquad \langle K(\boldsymbol{\xi}) \rangle_D = \frac{1}{|D|} \sum_{\boldsymbol{\xi} \in D} K(\boldsymbol{\xi}) \qquad \langle K(\boldsymbol{\xi}) \rangle_{\tilde{D}} = \frac{1}{|\tilde{D}|} \sum_{\boldsymbol{\xi} \in \tilde{D}} K(\boldsymbol{\xi})$$

The average $\langle K(\boldsymbol{\xi}) \rangle_{\tilde{D}}$ will in general depend on the microscopic realisation of the training set $\tilde{D}$. To quantify the goal and the progress of the student one finally defines an error $E[T(\boldsymbol{\xi}), S(\boldsymbol{\xi})] = E[T(\boldsymbol{\xi}), f[\boldsymbol{\xi}; \boldsymbol{J}]]$, which measures the mismatch between student answers and correct (teacher) answers for individual questions. The two key quantities of interest in supervised learning are the (time-dependent) averages of this error measure, calculated over the training set $\tilde{D}$ and the full question set D, respectively:

$$\begin{array}{ll} \text{Training Error :} & E_{\rm t}(\boldsymbol{J}) = \langle E[T(\boldsymbol{\xi}), f[\boldsymbol{\xi}; \boldsymbol{J}]] \rangle_{\tilde{D}} \\ \text{Generalization Error :} & E_{\rm g}(\boldsymbol{J}) = \langle E[T(\boldsymbol{\xi}), f[\boldsymbol{\xi}; \boldsymbol{J}]] \rangle_D \end{array} \tag{4.4}$$

These quantities are stochastic observables, since they are functions of the stochastically evolving vector $\boldsymbol{J}$. Their expectation values over the stochastic process (4.2) are given by

$$\begin{aligned} &\text{Mean Training Error :} && \langle E_{\rm t}\rangle = \langle\langle E[T(\boldsymbol{\xi}),f[\boldsymbol{\xi};\boldsymbol{J}]]\rangle\rangle_{\tilde{D}} \\ &\text{Mean Generalization Error :} && \langle E_{\rm g}\rangle = \langle\langle E[T(\boldsymbol{\xi}),f[\boldsymbol{\xi};\boldsymbol{J}]]\rangle\rangle_{D} \end{aligned} \tag{4.5}$$

Note that the prefix 'mean' refers to the stochasticity in the vector $\boldsymbol{J}$; both $\langle E_{\rm t}\rangle$ and $\langle E_{\rm g}\rangle$ will in general still depend on the realisation of the training set $\tilde{D}$.

The training error measures the performance of the student on the questions it could have been confronted with during the learning stage (in the case of on-line learning the student need not have seen all of them). The generalization error measures the student's performance on the full question set and its minimisation is therefore the main target of the process. The quality of a theory describing the dynamics of supervised learning can be measured by the degree to which it succeeds in predicting the values of $\langle E_{\rm t}\rangle$ and $\langle E_{\rm g}\rangle$ as a function of the iteration time t and for arbitrary choices made for the function $F[\ldots]$ that determines the details of the learning rules (4.1).

There are two main classes of situations in the supervised learning arena, which differ fundamentally in their dynamics and in the degree to which we can analyse them mathematically. The first class is the one where the training set $\tilde{D}$ is what we call 'complete': sufficiently large and sufficiently diverse to lead to a learning dynamics which in the limit $N\to\infty$ is identical to that of the situation where $\tilde{D}=D$. For example: in single perceptrons and in multi-layer perceptrons with a finite number of hidden nodes one finds, for the case where $D=\{-1,1\}^N$ and where the members of the training set $\tilde{D}$ are drawn at random from D, that completeness of the training set amounts to $\lim_{N\to\infty}N/|\tilde{D}|=0$. This makes sense: it means that for $N\to\infty$ there will be an infinite number of training examples per degree of freedom. For this class of models it is fair to say that the dynamics of learning can be fully analysed in a reasonably simple way. We will restrict ourselves to single perceptrons with various types of learning rules, since they form the most transparent playground for explaining how the mathematical techniques work. For multi-layer perceptrons with a finite number of hidden neurons and complete training sets the procedure to be followed is very similar[1]. The picture changes dramatically if we move away from complete training sets and consider those where the number of training examples is proportional to the number of degrees of freedom, i.e. in simple perceptrons and in two-layer perceptrons with a finite number of hidden neurons this implies $|\tilde{D}|=\alpha N$ $(0<\alpha<\infty)$. Now the dependence of the microscopic variables $\boldsymbol{J}$ on the realisation of the training set $\tilde{D}$ is non-negligible. However, if the questions in the training set are drawn at random from the full question set D one often finds that in the $N\to\infty$ limit the values of the *macroscopic* observables only depend on the size $|\tilde{D}|$ of the training set, not on its microscopic realisation. Here one needs much more powerful mathematical tools, which are as yet only partly available; therefore we will not deal with restricted training sets here.

[1] The situation is different if we try to deal with multi-layer perceptrons with a number of hidden neurons which scales linearly with the number of input channels N. This still poses an unsolved problem, even in the case of complete training sets.

4.2 Explicit Learning Rules

We will now derive explicitly macroscopic dynamical equations that describe the evolution in time for the error in large perceptrons, trained with several on-line learning rules to perform linearly separable tasks. In this section we restrict ourselves to complete training sets $\tilde{D} = D = \{-1,1\}^N$. There is consequently no difference between training and generalization error, and we can simply define $E = \lim_{N\to\infty}\langle E_{\rm g}\rangle = \lim_{N\to\infty}\langle E_{\rm t}\rangle$.

General On-Line Learning Rules. Consider a linearly separable binary classification task $T : \{-1,1\}^N \to \{-1,1\}$. It can be regarded as generated by a teacher perceptron with some unknown weight vector $\boldsymbol{B} \in \Re^N$, i.e. $T(\boldsymbol{\xi}) = \text{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})$, normalised according to $|\boldsymbol{B}| = 1$ (with the sign function $\text{sgn}(z>0) = 1$, $\text{sgn}(z<0) = -1$). A student perceptron with output $S(\boldsymbol{\xi}) = \text{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})$ (where $\boldsymbol{J} \in \Re^N$) is being trained in an on-line fashion using randomly drawn examples of input vectors $\boldsymbol{\xi} \in \{-1,1\}^N$ with corresponding teacher answers $T(\boldsymbol{\xi})$. The general picture of figure 4.1 thus specialises to figure 4.2. We exploit our knowledge of the perceptron's scaling properties and distinguish between the discrete time unit in terms of iteration steps, from now on to be denoted by $\mu = 1,2,3,\ldots$, and the scale-invariant time unit $t_\mu = \mu/N$. Our goal is to derive well-behaved differential equations in the limit $N\to\infty$, so we require weight changes occurring in intervals $\Delta t = \frac{1}{N}$ to be of order $\mathcal{O}(\frac{1}{N})$ as well. In terms of equation (4.1) this implies that $F[\ldots] = \mathcal{O}(\frac{1}{N})$. If, finally, we restrict ourselves to those rules where weight changes are made in the direction of the example vectors (which includes most popular rules), we obtain the generic[2] recipe

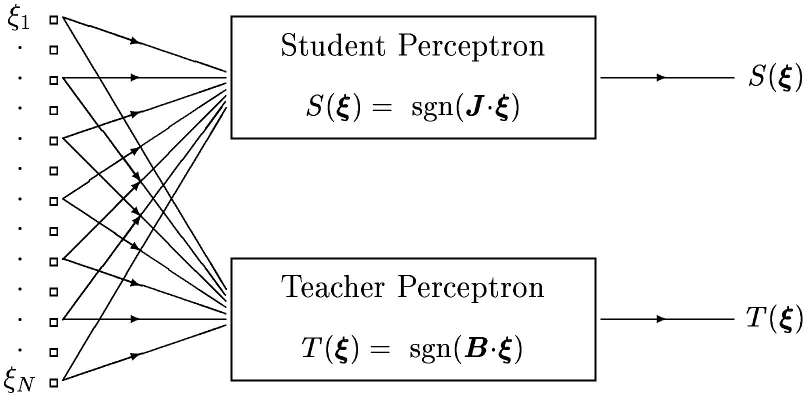

ξ_1
ξ_N
Student Perceptron
$S(\boldsymbol{\xi}) = \text{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})$
$S(\boldsymbol{\xi})$
Teacher Perceptron
$T(\boldsymbol{\xi}) = \text{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})$
$T(\boldsymbol{\xi})$

Figure 4.2: A student perceptron S is being trained according to on-line learning rules to perform a linearly separable operation, generated by some unknown teacher perceptron T.

$$\boldsymbol{J}(t_\mu + \frac{1}{N}) = \boldsymbol{J}(t_\mu) + \frac{1}{N}\eta(t_\mu)\boldsymbol{\xi}^\mu\ \text{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu] \tag{4.6}$$

Here $\eta(t_\mu)$ denotes a (possibly time-dependent) learning rate and $\boldsymbol{\xi}^\mu$ is the input vector selected at iteration step μ. $\mathcal{F}[\ldots]$ is an as yet arbitrary function of the length of the student weight vector and of the local fields u and v of student and teacher (note: $\mathcal{F}$ can depend on

[2] One can obviously write down more general rules, and also write the present recipe (4.6) in different ways.

the sign of the teacher field only, not on its magnitude). For example, for $\mathcal{F}[J;u,v]=1$ we obtain a Hebbian rule, for $\mathcal{F}[J;u,v]=\theta[-uv]$ we obtain the perceptron learning rule, etc.

We now try to solve the dynamics of the learning process in terms of the two macroscopic observables that play a special role in the perceptron convergence proof:

$$Q[\boldsymbol{J}]=\boldsymbol{J}^2 \qquad\qquad R[\boldsymbol{J}]=\boldsymbol{J}\cdot\boldsymbol{B} \tag{4.7}$$

(at this stage the selection of observables is still no more than intuition-driven guesswork). The formal approach would now be to derive an expression for the (time-dependent) probability density $P(Q,R)=\langle\delta[Q-Q[\boldsymbol{J}]]\delta[R-R[\boldsymbol{J}]]\rangle$, however, it turns out that in the present case[3] there is a short-cut. Squaring (4.6) and taking the inner product of (4.6) with the teacher vector $\boldsymbol{B}$ gives, respectively

$$\begin{aligned}
Q[\boldsymbol{J}(t_\mu+\frac{1}{N})] &= Q[\boldsymbol{J}(t_\mu)]+\frac{2}{N}\eta(t_\mu)(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\\
&\qquad +\frac{1}{N}\eta^2(t_\mu)\mathcal{F}^2[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\\
R[\boldsymbol{J}(t_\mu+\frac{1}{N})] &= R[\boldsymbol{J}(t_\mu)]+\frac{1}{N}\eta(t_\mu)|\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu|\mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]
\end{aligned}$$

(note: $\boldsymbol{\xi}^\mu\cdot\boldsymbol{\xi}^\mu=N$). After ℓ discrete update steps we will have accumulated ℓ such modifications, and will thus arrive at:

$$\frac{Q[\boldsymbol{J}(t_\mu+\ell/N)]-Q[\boldsymbol{J}(t_\mu)]}{\ell/N}=$$

$$\frac{1}{\ell}\sum_{m=0}^{\ell-1}\Big\{2\eta(t_\mu+\frac{m}{N})(\boldsymbol{J}(t_\mu+\frac{m}{N})\cdot\boldsymbol{\xi}^{\mu+m})\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m})\mathcal{F}[|\boldsymbol{J}(t_\mu+\frac{m}{N})|;\boldsymbol{J}(t_\mu+\frac{m}{N})\cdot\boldsymbol{\xi}^{\mu+m},\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}]$$
$$+\eta^2(t_\mu+\frac{m}{N})\mathcal{F}^2[|\boldsymbol{J}(t_\mu+\frac{m}{N})|;\boldsymbol{J}(t_\mu+\frac{m}{N})\cdot\boldsymbol{\xi}^{\mu+m},\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}]\Big\}$$

$$\frac{R[\boldsymbol{J}(t_\mu+\ell/N)]-R[\boldsymbol{J}(t_\mu)]}{\ell/N}=$$

$$\frac{1}{\ell}\sum_{m=0}^{\ell-1}\Big\{\eta(t_\mu+\frac{m}{N})|\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}|\mathcal{F}[|\boldsymbol{J}(t_\mu+\frac{m}{N})|;\boldsymbol{J}(t_\mu+\frac{m}{N})\cdot\boldsymbol{\xi}^{\mu+m},\boldsymbol{B}\cdot\boldsymbol{\xi}^{\mu+m}]\Big\}$$

All is still exact, but at this stage we will have to make an assumption which is not entirely satisfactory[4]. We assume that $\boldsymbol{J}(t_\mu+\frac{m}{N})\cdot\boldsymbol{\xi}^{\mu+m}\to\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^{\mu+m}$ if $N\to\infty$ for finite m. This is only true in a probabilistic sense, since, although $J_i(t_\mu+\frac{m}{N})=J_i(t_\mu)+\mathcal{O}(\frac{m}{N})$, the inner product is a sum of N terms. If for now, however, we accept this step and also choose learning rates which vary sufficiently slowly over time to guarantee existence of the limit $\lim_{N\to\infty}\eta(t_\mu)$, we find that by taking the limit $N\to\infty$, followed by the limit $\ell\to\infty$, three pleasant simplifications occur: (i) the time unit $t_\mu=\mu/N$ becomes a continuous variable, (ii) the left-hand sides of the above equations for the evolution of the observables Q and R become

[3] This will be different in the case of incomplete training sets.

[4] We will later find out that a more careful analysis gives the same results.

temporal derivatives, and (iii) the summations in the right-hand sides of these equations become averages of the training set. Upon putting $Q(t) = Q[\boldsymbol{J}(t)]$ and $R(t) = R[\boldsymbol{J}(t)]$ the result can be written as:

$$\frac{d}{dt}Q(t) = 2\eta(t)\langle(\boldsymbol{J}(t)\cdot\boldsymbol{\xi})\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi})\mathcal{F}[Q^{\frac{1}{2}}(t);\boldsymbol{J}(t)\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}} + \eta^2(t)\langle\mathcal{F}^2[Q^{\frac{1}{2}}(t);\boldsymbol{J}(t)\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}$$

$$\frac{d}{dt}R(t) = \eta(t)\langle|\boldsymbol{B}\cdot\boldsymbol{\xi}|\mathcal{F}[Q^{\frac{1}{2}}(t);\boldsymbol{J}(t)\cdot\boldsymbol{\xi},\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}}$$

The only dependence of the right-hand sides of these expressions on the microscopic variables $\boldsymbol{J}$ is via the student fields $\boldsymbol{J}(t)\cdot\boldsymbol{\xi} = Q^{\frac{1}{2}}(t)\hat{\boldsymbol{J}}(t)\cdot\boldsymbol{\xi}$, with $\hat{\boldsymbol{J}} = \boldsymbol{J}/|\boldsymbol{J}|$[5]. We therefore define the stochastic variables $x = \hat{\boldsymbol{J}}\cdot\boldsymbol{\xi}$ and $y = \boldsymbol{B}\cdot\boldsymbol{\xi}$ and their joint probability distribution $P_t(x,y)$:

$$P_t(x,y) = \langle\delta[x-\hat{\boldsymbol{J}}(t)\cdot\boldsymbol{\xi}]\delta[y-\boldsymbol{B}\cdot\boldsymbol{\xi}]\rangle_{\tilde{D}} \qquad \langle f(x,y)\rangle = \int dxdy\ P_t(x,y)f(x,y) \tag{4.8}$$

Using brackets without subscripts for joint field averages cannot cause confusion, since such expressions always *replace* averages over $\boldsymbol{J}$, rather than occur simultaneously. Our previous result now takes the form

$$\frac{d}{dt}Q(t) = 2\eta(t)Q^{\frac{1}{2}}(t)\langle x\ \mathrm{sgn}(y)\mathcal{F}[Q^{\frac{1}{2}}(t);Q^{\frac{1}{2}}(t)x,y]\rangle + \eta^2(t)\langle\mathcal{F}^2[Q^{\frac{1}{2}}(t);Q^{\frac{1}{2}}(t)x,y]\rangle \tag{4.9}$$

$$\frac{d}{dt}R(t) = \eta(t)\langle|y|\mathcal{F}[Q^{\frac{1}{2}}(t);Q^{\frac{1}{2}}(t)x,y]\rangle \tag{4.10}$$

Since the operation performed by the student does not depend on the length $|\boldsymbol{J}|$ of its weight vector, and since both Q and R involve $|\boldsymbol{J}|$, it will be convenient at this stage to switch to another (equivalent) pair of observables:

$$J(t) = |\boldsymbol{J}(t)| \qquad \omega(t) = \boldsymbol{B}\cdot\hat{\boldsymbol{J}}(t) \tag{4.11}$$

Using the relations $\frac{d}{dt}Q = 2J\frac{d}{dt}J$ and $\frac{d}{dt}R = J\frac{d}{dt}\omega + \omega\frac{d}{dt}J$, and upon dropping the various explicit time arguments (for notational convenience) we then find the compact expressions

$$\frac{d}{dt}J \ = \ \eta\langle x\ \mathrm{sgn}(y)\mathcal{F}[J;Jx,y]\rangle + \frac{\eta^2}{2J}\langle\mathcal{F}^2[J;Jx,y]\rangle \tag{4.12}$$

$$\frac{d}{dt}\omega \ = \ \frac{\eta}{J}\langle[|y|-\omega\,x\ \mathrm{sgn}(y)]\ \mathcal{F}[J;Jx,y]\rangle - \frac{\omega\eta^2}{2J^2}\langle\mathcal{F}^2[J;Jx,y]\rangle \tag{4.13}$$

Unless we manage to express $P(x,y)$ in terms of the pair (J,ω), however, the equations (4.12,4.13) do not constitute a solution of our problem , since we would still be forced to solve the original microsopic dynamical equations in order to find $P(x,y)$ as a function of time and work out (4.12,4.13).

The final stage of the argument is to assume that the joint probability distribution (4.8) has a Gaussian shape, since $\tilde{D} = \{-1,1\}^N$ and since all $\boldsymbol{\xi}\in\tilde{D}$ contribute equally to the average in (4.8). This will be true in the vast majority of cases, e.g. it is true with probability one if the vectors $\boldsymbol{J}$ and $\boldsymbol{B}$ are drawn at random from compact sets like $[-1,1]^N$, due to the

[5] This property of course depends crucially on our choice (4.6) made for the form of the learning rules.

central limit theorem[6]. Gaussian distributions are fully specified by their first and second order moments, which are here calculated trivially using $\langle\xi_i\rangle = 0$ and $\langle\xi_i\xi_j\rangle = \delta_{ij}$:

$$\langle x\rangle = \sum_i \hat{J}_i\langle\xi_i\rangle = 0 \qquad \langle y\rangle = \sum_i B_i\langle\xi_i\rangle = 0$$

$$\langle x^2\rangle = \sum_{ij} \hat{J}_i\hat{J}_j\langle\xi_i\xi_j\rangle = 1 \qquad \langle y^2\rangle = \sum_{ij} B_iB_j\langle\xi_i\xi_j\rangle = 1 \qquad \langle xy\rangle = \sum_{ij} \hat{J}_iB_j\langle\xi_i\xi_j\rangle = \omega$$

giving

$$P(x,y) = \frac{e^{-\frac{1}{2}[x^2+y^2-2xy\omega]/(1-\omega^2)}}{2\pi\sqrt{1-\omega^2}} \tag{4.14}$$

Note that $P(x,y) = P(y,x)$. The simple fact that $P(x,y)$ depends on time only through ω ensures that the two equations (4.12,4.13) are a *closed* set. Note also that now (4.12,4.13) are deterministic equations; apparently the fluctuations in the macroscopic observables $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ vanish in the $N \to \infty$ limit.

Finally, the generalization error $E_{\rm g}$ (here identical to the training error $E_{\rm t}$ due to $\tilde{D} = D$) can be expressed in terms of our macroscopic observables. We define the error made in a single classification of an input $\boldsymbol{\xi}$ as $E[T(\boldsymbol{\xi}), S(\boldsymbol{\xi})] = \theta[-(\boldsymbol{B}\cdot\boldsymbol{\xi})(\boldsymbol{J}\cdot\boldsymbol{\xi})] \in \{0,1\}$. Averaged over D this gives the probability of a misclassification for randomly drawn questions $\boldsymbol{\xi} \in D$:

$$\lim_{N\to\infty} E_{\rm g}(\boldsymbol{J}(t)) = \lim_{N\to\infty} \langle[\theta[-(\boldsymbol{B}\cdot\boldsymbol{\xi})(\boldsymbol{J}(t)\cdot\boldsymbol{\xi})]\rangle_D = \langle\theta[-xy]\rangle$$

$$= \int_0^\infty\int_0^\infty dxdy\,[P(x,-y) + P(-x,y)]$$

The generalization error (from this stage onwards to be denoted simply by E) also evolves deterministically for $N \to \infty$, and can be expressed purely in terms of the observable ω. The integral (with the distribution (4.14)) can even be done analytically (see appendix) and produces the simple result

$$E = \frac{1}{\pi}\arccos(\omega) \tag{4.15}$$

The macrosopic equations (4.12,4.13) can now equivalently be written in terms of the pair (J, E). We have hereby achieved our goal: we have derived a closed set of deterministic equations for a small number (two) of macroscopic observables, valid for $N \to \infty$, and we know the generalization error at any time.

Hebbian Learning with Constant Learning Rate. We will now work out our general result (4.12,4.13,4.14) for specific members of the general class (4.6) of on-line learning rules. The simplest non-trivial choice to be made is the Hebbian rule, obtained by choosing $\mathcal{F}[J; Jx, y] = 1$, with a constant learning rate η:

$$\boldsymbol{J}(t_\mu + \frac{1}{N}) = \boldsymbol{J}(t_\mu) + \frac{\eta}{N}\boldsymbol{\xi}^\mu\ \text{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu) \tag{4.16}$$

[6] It is not true for all choices of $\mathbf{J}$ and $\mathbf{B}$. A trivial counter-example is $J_k = \delta_{k1}$, less trivial counter-examples are e.g. $J_k = e^{-k}$ and $J_k = k^{-\gamma}$ with $\gamma > \frac{1}{2}$.

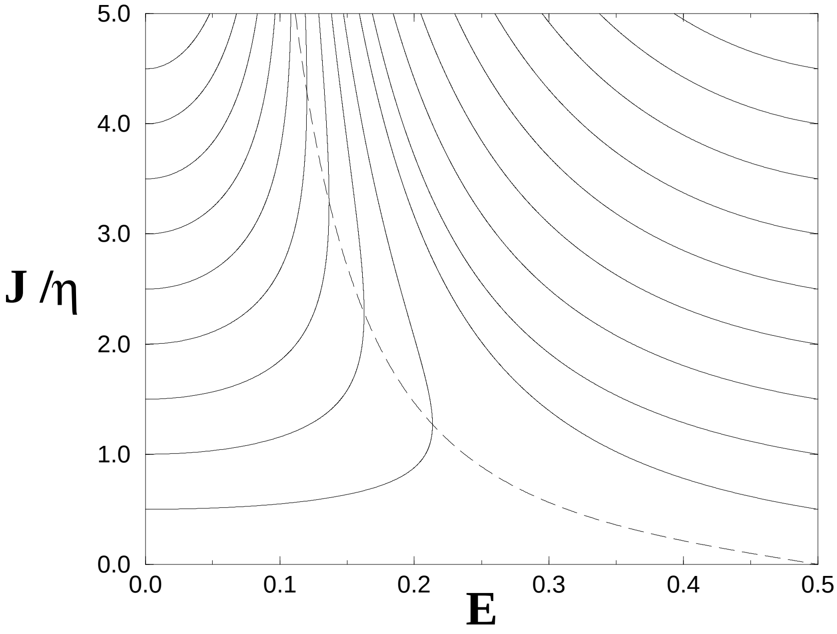


Figure 4.3: Flow in the (E, J) plane generated by the Hebbian learning rule with constant learning rate η, in the limit $N \to \infty$. Dashed: the line where $dE/dt = 0$ ($dJ/dt > 0$ for any (E, J)). Note that the flow asymptotically gives $E \to 0$ and $J \to \infty$.

Equations (4.12) and (4.13), describing the macroscopic dynamics generated by (4.16) in the limit $N \to \infty$ now become

$$\frac{d}{dt}J = \eta\langle x \ \mathrm{sgn}(y)\rangle + \frac{\eta^2}{2J} \qquad\qquad \frac{d}{dt}\omega = \frac{\eta}{J}\langle |y| - \omega x \ \mathrm{sgn}(y)\rangle - \frac{\omega\eta^2}{2J^2}$$

or, in more explicit form with the function $P(x, y)$ (4.14):

$$\begin{aligned}\frac{d}{dt}J &= \eta\int\!\!\int dx dy\, x \ \mathrm{sgn}(y)P(x,y) + \frac{\eta^2}{2J} \\ \frac{d}{dt}\omega &= \frac{\eta}{J}\int\!\!\int dx dy\, |y|P(x,y) - \frac{\omega\eta}{J}\int\!\!\int dx dy\, x \ \mathrm{sgn}(y)P(x,y) - \frac{\omega\eta^2}{2J^2}\end{aligned}$$

The integrals in these equations can be calculated analytically (see appendix) and we get

$$\frac{d}{dt}J = \omega\eta\sqrt{\frac{2}{\pi}} + \frac{\eta^2}{2J} \qquad\qquad \frac{d}{dt}\omega = (1-\omega^2)\frac{\eta}{J}\sqrt{\frac{2}{\pi}} - \frac{\omega\eta^2}{2J^2}$$

Thus, upon elimination of the observable ω using equation (4.15), we arrive at the following closed differential equations in terms of J and E:

$$\frac{d}{dt}J = \eta\cos(\pi E)\sqrt{\frac{2}{\pi}} + \frac{\eta^2}{2J} \tag{4.17}$$

$$\frac{d}{dt}E = -\frac{\eta\sin(\pi E)}{\pi J}\sqrt{\frac{2}{\pi}} + \frac{\eta^2}{2\pi J^2\tan(\pi E)} \tag{4.18}$$

The flow in the (E, J) plane described by these equations is drawn in figure 4.3 (which is obtained by numerical solution of (4.17,4.18)). From (4.17) it follows that $\frac{d}{dt}J > 0 \; \forall t \geq 0$. From (4.18) it follows that $\frac{d}{dt}E = 0$ along the line

$$J_c(E) = \frac{\eta \cos(\pi E)}{2\sin^2(\pi E)}\sqrt{\frac{\pi}{2}}$$

(drawn as a dashed line in figure 4.3).

Let us now investigate the temporal properties of the solution (4.17,4.18), and work out their predictions for the asymptotic decay of the generalization error. For small values of E equations (4.17,4.18) yield

$$\frac{d}{dt}J = \eta\sqrt{\frac{2}{\pi}} + \frac{\eta^2}{2J} + \mathcal{O}(E^2) \tag{4.19}$$

$$\frac{d}{dt}E = -\frac{\eta E}{J}\sqrt{\frac{2}{\pi}} + \frac{\eta^2}{2\pi^2 J^2 E} + \mathcal{O}(E^3/J, E/J^2) \tag{4.20}$$

From (4.19) we infer that $J \sim \eta t\sqrt{\frac{2}{\pi}}$ for $t \to \infty$. Subsitution of this asymptotic solution into equation (4.20) gives

$$\frac{d}{dt}E = -\frac{E}{t} + \frac{1}{4\pi E t^2} + \mathcal{O}(E^3/t, E/t^2) \qquad (t \to \infty) \tag{4.21}$$

We insert the ansatz $E = At^{-\alpha}$ into equation (4.21) and get the solution $A = 1/\sqrt{2\pi}$, $\alpha = 1/2$. This implies that (in the $N \to \infty$ limit) on-line Hebbian learning with complete training sets produces an asymptotic decay of the generalization of the form

$$E \sim \frac{1}{\sqrt{2\pi t}} \qquad (t \to \infty) \tag{4.22}$$

Figures 4.6, 4.7 and 4.8 will show the theoretical results of this section together with the results of doing numerical simulations of the learning rule (4.16) and with similar results for other on-line learning rules with constant learning rates. The agreement between theory and simulations is quite convincing.

Perceptron Learning with Constant Learning Rate. Our second application of (4.12,4.13,4.14) is making the choice $\mathcal{F}[J; Jx, y] = \theta[-xy]$ in equation (4.6), with constant learning rate η, which produces the perceptron learning algorithm:

$$\boldsymbol{J}(t_\mu + \frac{1}{N}) = \boldsymbol{J}(t_\mu) + \frac{\eta}{N}\boldsymbol{\xi}^\mu \theta\left[-(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\right] \tag{4.23}$$

In other words: the student weights are updated in accordance with the Hebbian rule only when $\text{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}) = -\,\text{sgn}(\boldsymbol{J}\cdot\boldsymbol{\xi})$, i.e. when student and teacher are not in agreement. Equations (4.12,4.13) now become

$$\begin{aligned}\frac{d}{dt}J &= \eta\langle x \;\text{sgn}(y)\theta[-xy]\rangle + \frac{\eta^2}{2J}\langle\theta[-xy]\rangle \\ &= \eta \int\!\!\int dxdy \; x \;\text{sgn}(y)\theta[-xy]P(x,y) + \frac{\eta^2}{2J}\int\!\!\int dxdy \; \theta[-xy]P(x,y)\end{aligned}$$

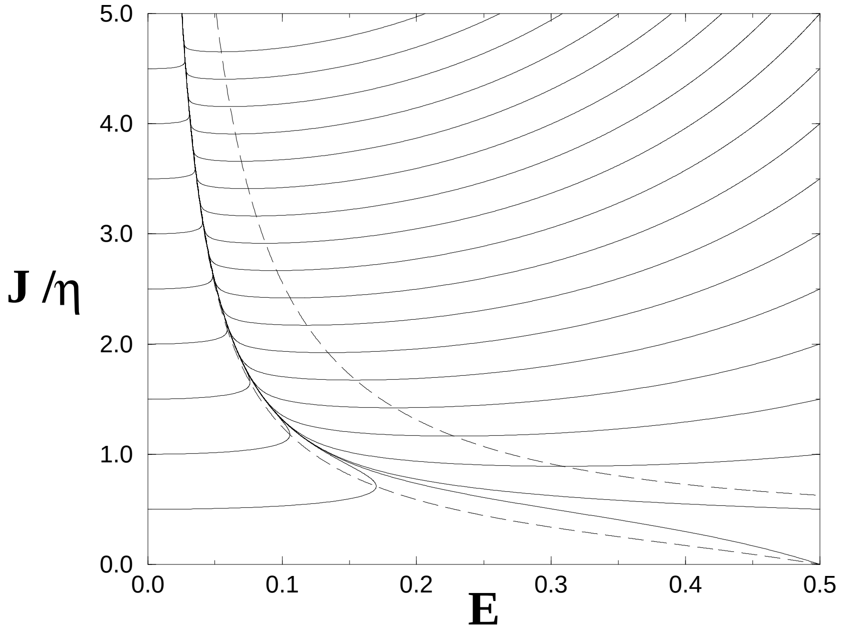


Figure 4.4: Flow in the (E, J) plane generated by the perceptron learning rule with constant learning rate η, in the limit $N \to \infty$. Dashed: the two lines where $dE/dt = 0$ and $dJ/dt = 0$, respectively. Note that the flow is attracted into the gully between these two dashed lines and asymptotically gives $E \to 0$ and $J \to \infty$.

$$\begin{aligned}
\frac{d}{dt}\omega &= \frac{\eta}{J}\langle[|y| - \omega x \text{ sgn}(y)]\,\theta[-xy]\rangle - \frac{\omega\eta^2}{2J^2}\langle\theta[-xy]\rangle \\
&= \frac{\eta}{J}\int\!\!\int dxdy\ |y|\theta[-xy]P(x,y) - \frac{\omega\eta}{J}\int\!\!\int dxdy\ x\ \text{sgn}(y)\theta[-xy]P(x,y) \\
&\quad -\frac{\omega\eta^2}{2J^2}\int\!\!\int dxdy\ \theta[-xy]P(x,y)
\end{aligned}$$

with $P(x, y)$ given by (4.14). As before the various Gaussian integrals occurring in these expressions can be done analytically (see appendix), which results in

$$\frac{d}{dt}J = -\frac{\eta(1-\omega)}{\sqrt{2\pi}} + \frac{\eta^2}{2\pi J}\arccos(\omega) \qquad \frac{d}{dt}\omega = \frac{\eta(1-\omega^2)}{\sqrt{2\pi}J} - \frac{\omega\eta^2}{2\pi J^2}\arccos(\omega)$$

Elimination of ω using (4.15) then gives us the dynamical equations in terms of the pair (J, E):

$$\frac{d}{dt}J = -\frac{\eta(1-\cos(\pi E))}{\sqrt{2\pi}} + \frac{\eta^2 E}{2J} \tag{4.24}$$

$$\frac{d}{dt}E = -\frac{\eta\sin(\pi E)}{\pi\sqrt{2\pi}J} + \frac{\eta^2 E}{2\pi J^2\tan(\pi E)} \tag{4.25}$$

Figure 4.4 shows the flow in the (E, J) plane, obtained by numerical solution of (4.24,4.25).

The two lines where $\frac{d}{dt}J = 0$ and where $\frac{d}{dt}E = 0$ are found to be $J_{c,1}(E)$ and $J_{c,2}(E)$, respectively:

$$J_{c,1}(E) = \eta\sqrt{\frac{\pi}{2}}\frac{E}{1-\cos(\pi E)} \qquad J_{c,2}(E) = \eta\sqrt{\frac{\pi}{2}}\frac{E\cos(\pi E)}{1-\cos^2(\pi E)}$$

For $E \in [0, 1/2]$ one always has $J_{c,1}(E) \geq J_{c,2}(E)$, with equality only if $(J, E) = (\infty, 0)$. Figure 4.4 shows that the flow is drawn into the gully between the curves $J_{c,1}(E)$ and $J_{c,2}(E)$.

As with the Hebbian rule we now wish to investigate the asymptotic behaviour of the generalization error. To do this we expand equations (4.24,4.25) for small E:

$$\frac{d}{dt}J = -\frac{\eta\pi^2E^2}{2\sqrt{2\pi}} + \frac{\eta^2 E}{2J} + \mathcal{O}(E^4)$$

$$\frac{d}{dt}E = -\frac{\eta E}{\sqrt{2\pi}J} + \frac{\eta^2}{2\pi^2J^2} - \frac{\eta^2E^2}{6J^2} + \mathcal{O}(E^3)$$

For small E and large t we know that $J \sim J_{c,1}(E) \sim 1/E$. Making the ansatz $J = A/E$ (and hence $\frac{d}{dt}E = -\frac{E^2}{A}\frac{d}{dt}J$) leads to a situation where we have two equivalent differential equations for E:

$$\frac{d}{dt}E = \frac{\eta\pi^2E^4}{2\sqrt{2\pi}A} - \frac{\eta^2E^4}{2A^2} + \mathcal{O}(E^6)$$

$$\frac{d}{dt}E = -\frac{\eta E^2}{\sqrt{2\pi}A} + \frac{\eta^2E^2}{2\pi^2A^2} + \mathcal{O}(E^4)$$

Since both describe the same dynamics, the leading term of the second expression should be identical to that of the first, i.e. $\mathcal{O}(E^4)$, giving us the condition $A = \frac{\eta\sqrt{2\pi}}{2\pi^2}$. Substitution of this condition into the first expression for $\frac{d}{dt}E$ then gives

$$\frac{d}{dt}E = -\frac{1}{2}\pi^3E^4 + \mathcal{O}(E^5) \qquad (t \to \infty)$$

which has the solution

$$E \sim \left(\frac{2}{3}\right)^{1/3}\pi^{-1}t^{-1/3} \qquad (t \to \infty) \tag{4.26}$$

We find, somewhat surprisingly, that in large systems ($N \to \infty$) the on-line perceptron learning rule is asymptotically much slower in converging towards the desired $E = 0$ state than the simpler Hebbian rule. This will be different if we allow for time-dependent learning rates. Figures 4.6, 4.7 and 4.8 will show the theoretical results on the perceptron rule together with the results of doing numerical simulations and together with similar results for other on-line learning rules. Again the agreement between theory and experiment is quite satisfactory.

AdaTron Learning with Constant Learning Rate. As our third application we analyse the macroscopic dynamics of the AdaTron learning rule, corresponding to the choice $\mathcal{F}[J; Jx, y] = |Jx|\theta[-xy]$ in the general recipe (4.6). As in the perceptron rule, modifications are made only when student and teacher are in disagreement; however, here the modification made is proportional to the magnitude of the student's local field. Students are punished in proportion to their confidence in the wrong answer. The rationale is that wrong student answers

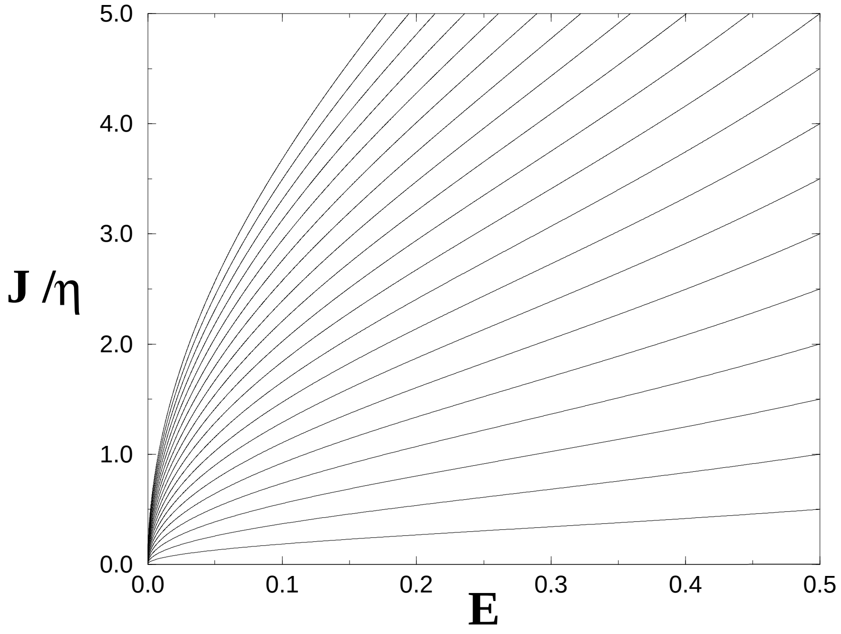


Figure 4.5: Flow in the (E, J) plane generated by the AdaTron learning rule with constant learning rate $\eta = 1$, in the limit $N \to \infty$ (in this case the influence of the value of the learning rate on the flow is more than just a rescaling of the length J).

$S(\boldsymbol{\xi}) =$ sgn$(\boldsymbol{J}\cdot\boldsymbol{\xi})$ with large values of $|\boldsymbol{J}\cdot\boldsymbol{\xi}|$ require more rigorous corrections to $\boldsymbol{J}$ to be remedied than those with small values of $|\boldsymbol{J}\cdot\boldsymbol{\xi}|$.

$$\boldsymbol{J}(t_\mu+\frac{1}{N}) = \boldsymbol{J}(t_\mu) + \frac{\eta}{N}\boldsymbol{\xi}^\mu\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)|\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu|\theta[-(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu)] \tag{4.27}$$

Working out the general equations (4.12,4.13) for the learning rule (4.27) gives

$$\begin{aligned}
\frac{d}{dt}J &= \eta J\int\!\!\int dxdy\ x|x|\ \mathrm{sgn}(y)\theta[-xy]P(x,y) + \frac{1}{2}\eta^2 J\int\!\!\int dxdy\ x^2\theta[-xy]P(x,y)\\
\frac{d}{dt}\omega &= \eta\int\!\!\int dxdy\ |xy|\theta[-xy]P(x,y) - \eta\omega\int\!\!\int dxdy\ x|x|\ \mathrm{sgn}(y)\theta[-xy]P(x,y)\\
&\qquad -\frac{1}{2}\omega\eta^2\int\!\!\int dxdy\ x^2\theta[-xy]P(x,y)
\end{aligned}$$

All integrals can again be done analytically (see appendix), so that we obtain explicit macroscopic flow equations:

$$\frac{d}{dt}J = \frac{J}{\omega}[\eta-\frac{\eta^2}{2}]I_2(\omega) \qquad\qquad \frac{d}{dt}\omega = \eta I_1(\omega) - [\eta-\frac{\eta^2}{2}]I_2(\omega)$$

with the short-hands

$$I_1(\omega) = \frac{(1-\omega^2)^{3/2}}{\pi} - \frac{\omega(1-\omega^2)}{\pi}\arccos(\omega) + \frac{\omega^2\sqrt{1-\omega^2}}{\pi} - \frac{\omega^3}{\pi}\arccos(\omega)$$

$$I_2(\omega) = -\frac{\omega(1-\omega^2)}{\pi}\arccos(\omega) + \frac{\omega^2\sqrt{1-\omega^2}}{\pi} - \frac{\omega^3}{\pi}\arccos(\omega)$$

The usual translation from equations for the pair (J, ω) into one involving the pair (J, E), following (4.15), turns out to simplify matters considerably, since it gives

$$\frac{d}{dt}J = J[\frac{\eta^2}{2} - \eta]\left[E - \frac{\cos(\pi E)\sin(\pi E)}{\pi}\right] \tag{4.28}$$

$$\frac{d}{dt}E = -\frac{\eta\sin^2(\pi E)}{\pi^2} + \frac{\eta^2 E}{2\pi\tan(\pi E)} - \frac{\eta^2\cos^2(\pi E)}{2\pi^2} \tag{4.29}$$

The flow described by the equations (4.28,4.29) is shown in figure 4.5, for the case $\eta = 1$. In contrast with the Hebbian and the perceptron learning rules we here observe from the equations (4.28,4.29) that the learning rate η cannot be eliminated from the macroscopic laws by a rescaling of the weight vector length J. Moreover, the state $E = 0$ is stable only for $\eta < 3$, in which case $\frac{d}{dt}E < 0$ for all t. For $\eta < 2$ one has $\frac{d}{dt}J < 0$ for all t, for $\eta = 2$ one has $J(t) = J(0)$ for all t, and for $2 < \eta < 3$ we have $\frac{d}{dt}J > 0$ for all t.

For small E equation (4.29) reduces to

$$\frac{d}{dt}E = [\frac{\eta^2}{3} - \eta]E^2 + \mathcal{O}(E^4)$$

giving

$$E \sim \frac{3t^{-1}}{\eta(3-\eta)} \qquad (t \to \infty) \tag{4.30}$$

For $\eta = 1$, which gives the standard representation of the AdaTron alrorithm, we find $E \sim \frac{3}{2}t^{-1}$. Note from equation (4.28) that for the AdaTron rule there is a value for η which normalises the length J of the student's weight vector, $\eta = 2$, which again gives $E \sim \frac{3}{2}t^{-1}$. The optimal value for η, however, is $\eta = \frac{3}{2}$ in which case we find $E \sim \frac{4}{3}t^{-1}$ (see (4.30)).

Theory Versus Simulations. We close this section with results of comparing the dynamics described by the various macroscopic flow equations with the results of measuring the error E during numerical simulations of the various (microscopic) learning rules discussed so far. This will serve to support the analysis and its implicit and explicit assumptions, but also illustrates how the three learning rules compare among one another. Figures 4.6 and 4.7 show the initial stage of the learning processes, for initialisations corresponding to random guessing ($E = 0.5$) and almost correct classification (E small), respectively (note that for the perceptron and Adatron rules starting at precisely $E = 0$ produces in finite systems a stationary state). Here the solutions of the flow equations (solid lines) were obtained by numerical iteration. The initial increase in the error E, as observed for the Hebbian and perceptron rule, following initialisation with small values of E can be understood as follows. The error depends only on the angle of the weight vector $\boldsymbol{J}$, not on its length J, this means that the modifications generated by the Hebbian and perceptron learning rules (which are of uniform magnitude) generate large changes in E when J is small, but small changes in E when J is large, with corresponding effects on the stability of low E states. The AdaTron rule, in contrast, involves weight changes which scale with the length J, so that the stability of the $E = 0$ state does not depend on the value of J. Figure 4.8 shows the asymptotic relaxation of the error E, in a log-log plot, together with the three corresponding asymptotic

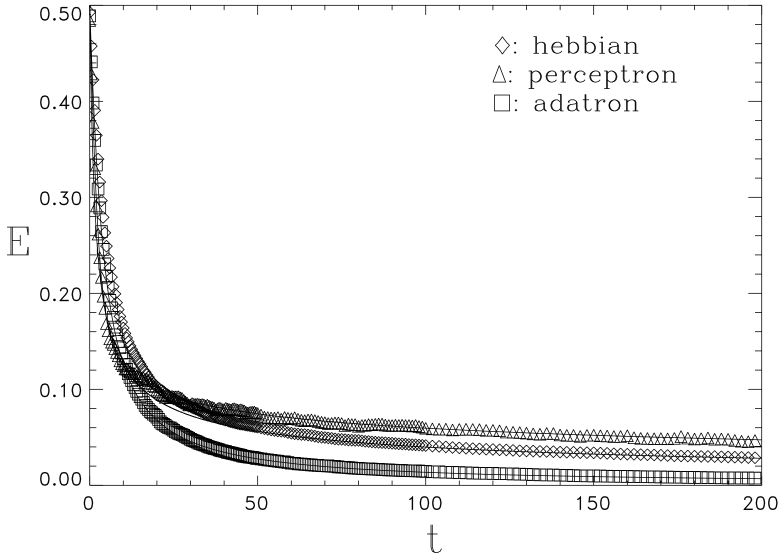


Figure 4.6: Evolution in time of the generalization error E as measured during numerical simulations (with $N = 1000$ neurons) of three different learning rules: Hebbian (diamonds), perceptron (triangles) and AdaTron (squares). Initial state: $E(0) = \frac{1}{2}$ (random guessing) and $J(0) = 1$. Learning rate: $\eta = 1$. The solid lines give for each learning rule the prediction of the $N = \infty$ theory, obtained by numerical solution of the flow equations for (E, J).

(power law) predictions (4.22,4.26,4.30). All simulations were carried out with networks of $N = 1000$ neurons, which apparently is already sufficiently large for the $N = \infty$ theory to apply. The teacher weight vectors $\boldsymbol{B}$ were in all cases drawn at random from $[-1,1]^N$. We conclude that the theory describes the simulations essentially perfectly.

4.3 Optimised Learning Rules

We now set out to use our macroscopic equations in 'reverse mode'. Rather than calculate the macroscopic dynamics for a given choice of learning rule, we will try to find learning rules that optimise the macroscopic dynamical laws in the sense that they produce the fastest decay towards the desired $E = 0$ state. As a bonus it will turn out that in many cases we can even solve the corresponding macroscopic differential equations analytically, and find explicit expressions for $E(t)$, or rather its inverse $t(E)$.

Time-Dependent Learning Rates. First we illustrate how modifying existing learning rules in a simple way, by just allowing for suitably chosen time-dependent learning rates $\eta(t)$, can already lead to a drastic improvement in the asymptotic behaviour of the error E.

We will inspect two specific choices of time-dependent learning rates for the perceptron rule. Without loss of generality we can always put $\eta(t) = K(t)J(t)$ in our dynamic equations (for notational convenience we will drop the explicit time argument of K). This choice will enable us to decouple the dynamics of J from that of the generalization error E. For the

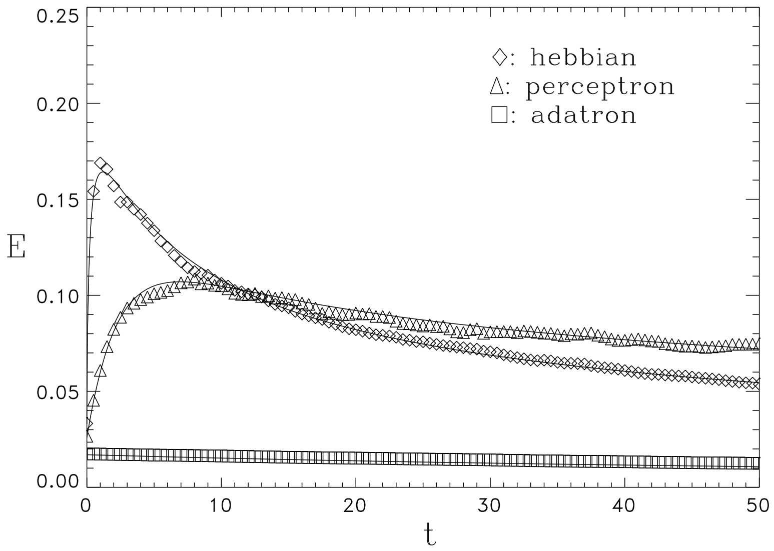


Figure 4.7: Evolution in time of the generalization error E as measured during numerical simulations (with $N = 1000$ neurons) of three different learning rules: Hebbian (diamonds), perceptron (triangles) and AdaTron (squares). Initial state: $E(0) \approx 0.025$ and $J(0) = 1$. Learning rate: $\eta = 1$. The solid lines give for each learning rule the prediction of the $N = \infty$ theory, obtained by numerical solution of the flow equations for (E, J).

perceptron rule we subsequently find equation (4.25) being replaced by

$$\frac{d}{dt}E = -\frac{K\sin(\pi E)}{\pi\sqrt{2\pi}} + \frac{K^2 E}{2\pi\tan(\pi E)}$$

giving for small E

$$\frac{d}{dt}E = -\frac{KE}{\sqrt{2\pi}} + \frac{K^2}{2\pi^2} + \mathcal{O}(K^2E^2)$$

In order to obtain $E \to 0$ for $t \to \infty$ it is clear that we need $K \to 0$. Applying the ansatz $E = A/t^\alpha$, $K = B/t^\beta$ for the asymptotic forms in the previous equation produces

$$-At^{-\alpha-1} = \frac{-ABt^{-\alpha-\beta}}{\sqrt{2\pi}} + \frac{B^2t^{-2\beta}}{2\pi^2} + \mathcal{O}(t^{-2\alpha-2\beta})$$

and so: $\alpha = \beta = 1$ and $A = \frac{1}{\pi\sqrt{2\pi}}\frac{B^2}{(B-\sqrt{2\pi})}$. Our aim is to obtain the fastest approach of the $E = 0$ state, i.e. we wish to maximise α (for which we found $\alpha = 1$) and subsequently minimise A. Apparently the value of B for which A is minimized is $B = 2\sqrt{2\pi}$, in which case we obtain the error decay given by

$$\eta \sim \frac{2J\sqrt{2\pi}}{t}: \qquad E \sim \frac{4}{\pi t} \qquad (t \to \infty) \tag{4.31}$$

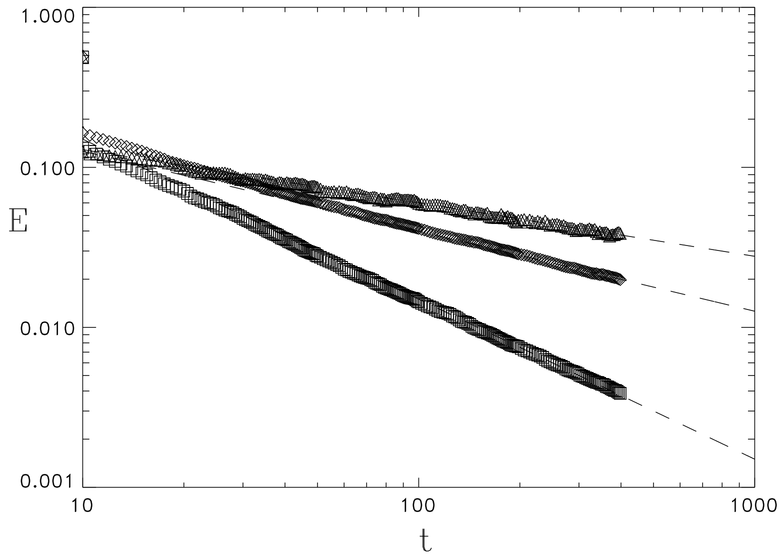


Figure 4.8: Asymptotic behaviour of the generalization error E measured during numerical simulations (with $N = 1000$) of three different learning rules: Hebbian (diamonds, middle curve), perceptron (triangles, upper curve) and AdaTron (squares, lower curve). Initial state: $E(0) = \frac{1}{2}$ and $J(0) = 1$. Learning rate: $\eta = 1$. The dashed lines give for each learning rule the corresponding power law predicted by the $N = \infty$ theory (equations (4.22,4.26,4.30), respectively).

This is clearly a great improvement upon the result for the perceptron rule with constant η, i.e. equation (4.26); in fact it is the fastest relaxation we have derived so far.

Let us now move to an alternative choice for the time-dependent learning rate for the perceptron. According to equation (4.24) there is one specific recipe for $\eta(t)$ such that the length J of the student's weight vector will remain constant, given by

$$\eta = \sqrt{\frac{2}{\pi}}\frac{J}{E}(1-\cos(\pi E)) \tag{4.32}$$

Making this choice converts equation (4.25) for the evolution of E into

$$\frac{d}{dt}E = -\frac{(1-\cos(\pi E))^2}{\pi^2 E \sin(\pi E)} \tag{4.33}$$

Equation (4.33) can be written in the form $\frac{d}{dE}t = g(E)$, so that $t(E)$ becomes a simple integral which can be done analytically, with the result

$$t(E) = \frac{\pi E + \sin(\pi E)}{1-\cos(\pi E)} - \frac{\pi E_0 + \sin(\pi E_0)}{1-\cos(\pi E_0)} \tag{4.34}$$

(which can also be verified directly by substitution into (4.33)). Expansion of (4.34) and (4.32) for small E gives the asymptotic behaviour also encountered in (4.31):

$$\eta \sim \frac{2J\sqrt{2\pi}}{t}, \qquad E \sim \frac{4}{\pi t} \qquad (t \to \infty) \tag{4.35}$$

It might appear that implementation of the recipe (4.32) is in practice impossible, since it involves information which is not available to the student perceptron (namely the instantaneous error E). However, since we know (4.34) we can simply calculate the required $\eta(t)$ explicitly as a function of time.

One has to be somewhat careful in extrapolating results such as those obtained in this section. For instance, choosing the time-dependent learning rate (4.32) enforces the constraint $\boldsymbol{J}^2(t)=1$ in the macroscopic equations for $N\to\infty$. This is not identical to choosing $\eta(t_\mu)$ in the original equation (4.6) such as to enforce $\boldsymbol{J}^2(t_\mu+\frac{1}{N})=\boldsymbol{J}^2(t_\mu)$ at the level of individual iteration steps, as can be seen by working out the dynamical laws. The latter case would correspond to the *microscopically fluctuating* choice

$$\eta(t_\mu)=-2\frac{(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)}{\mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]}\qquad\text{if}\quad \mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\neq 0$$

If we now choose for example $\mathcal{F}[J;Jx,y]=\theta[-xy]$, implying $\eta(t_\mu)=2|\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu|$, we find by insertion into (4.6) that the perceptron rule with 'hard' weight normalisation at each iteration step via adaptation of the learning rate is identical to the AdaTron rule with constant learning rate $\eta=2$. We know therefore that in this case one obtains $E\sim 3/2t$, whereas for the Perceptron rule with 'soft' weight normalisation via (4.32) (see the analysis above) one obtains $E\sim 4/\pi t$. Apparently the two procedures are not equivalent.

Spherical On-Line Learning Rules. We arrive in a natural way at the question of how to find the optimal time-dependent learning rate for any given learning rule, or more generally: of how to find the optimal learning rule. This involves variational calculations in two-dimensional flows (since our macroscopic equations are defined in terms of the evolving pair (J,E)). Such calculations would be much simpler if our macroscopic equations were just one-dimensional, e.g. describing only the evolution of the error E with a stationary (or simply irrelevant) value of the length J. Often it will turn out that for finding the optimal learning rate or the optimal learning rule the problem can indeed be reduced to a one-dimensional one. To be able to obtain results also for those cases where this reduction does not happen we will now construct so-called spherical learning rules, where $\boldsymbol{J}^2(t)=1$ for all t. This can be arranged in several equivalent ways.

The first method is to add to the general rule (4.6) a term proportional to the instantaneous weight vector $\boldsymbol{J}$, whose sole purpose is to achieve the constraint $\boldsymbol{J}^2=1$:

$$\boldsymbol{J}(t_\mu+\frac{1}{N})=\boldsymbol{J}(t_\mu)+\frac{1}{N}\left\{\eta(t_\mu)\boldsymbol{\xi}^\mu\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]-\lambda(t_\mu)\boldsymbol{J}(t_\mu)\right\}\quad(4.36)$$

The evolution of the two observables $Q[\boldsymbol{J}]$ and $R[\boldsymbol{J}]$ (4.7) is now given by

$$\begin{aligned}Q[\boldsymbol{J}(t_\mu+\frac{1}{N})] &= Q[\boldsymbol{J}(t_\mu)](1-\frac{2\lambda(t_\mu)}{N})+\frac{2}{N}\eta(t_\mu)(\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\\ &\qquad +\frac{1}{N}\eta^2(t_\mu)\mathcal{F}^2[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]+\mathcal{O}(N^{-2})\\ R[\boldsymbol{J}(t_\mu+\frac{1}{N})] &= R[\boldsymbol{J}(t_\mu)](1-\frac{\lambda(t_\mu)}{N})+\frac{1}{N}\eta(t_\mu)|\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu|\mathcal{F}[|\boldsymbol{J}(t_\mu)|;\boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu,\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\end{aligned}$$

Following the procedure of section 1.2 to arrive at the $N \to \infty$ limit of the dynamical equations for Q and R then leads to (we drop explicit time arguments for notational convenience):

$$\frac{d}{dt}Q = 2\eta Q^{\frac{1}{2}}\langle x\ \mathrm{sgn}(y)\mathcal{F}[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y]\rangle + \eta^2\langle \mathcal{F}^2[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y]\rangle - 2\lambda Q$$

$$\frac{d}{dt}R = \eta\langle |y|\mathcal{F}[Q^{\frac{1}{2}}; Q^{\frac{1}{2}}x, y]\rangle - \lambda R$$

We now choose the function $\lambda(t)$ such that $Q(t) = 1$ for all $t \geq 0$. This ensures that $R(t) = \omega(t) = \hat{\boldsymbol{J}}(t)\cdot\boldsymbol{B}$, and gives (via $\frac{d}{dt}Q = 0$) a recipe for $\lambda(t)$

$$\lambda = \eta\langle x\ \mathrm{sgn}(y)\mathcal{F}[1; x, y]\rangle + \frac{1}{2}\eta^2\langle \mathcal{F}^2[1; x, y]\rangle$$

which can then be substituted into our equation for $\frac{d}{dt}\omega$:

$$\frac{d}{dt}\omega = \eta\langle [|y| - \omega x\ \mathrm{sgn}(y)]\,\mathcal{F}[1; x, y] - \frac{1}{2}\omega\eta^2\langle \mathcal{F}^2[1; x, y]\rangle \tag{4.37}$$

with averages as usual defined with respect to the Gaussian joint field distribution (4.14), which depends only on ω, so that equation (4.37) is indeed autonomous.

The second method to arrange the constraint $\boldsymbol{J}^2 = 1$ is to explicitly normalise the weight vector $\boldsymbol{J}$ after each modification step, i.e.

$$\boldsymbol{J}(t_\mu + \frac{1}{N}) = \frac{\boldsymbol{J}(t_\mu) + \frac{1}{N}\eta(t_\mu)\boldsymbol{\xi}^\mu\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathcal{F}[1; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]}{|\boldsymbol{J}(t_\mu) + \frac{1}{N}\eta(t_\mu)\boldsymbol{\xi}^\mu\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathcal{F}[1; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]|} \tag{4.38}$$

$$= \hat{\boldsymbol{J}}(t_\mu) + \frac{1}{N}\eta(t_\mu)\left\{\left[\boldsymbol{\xi}^\mu - \hat{\boldsymbol{J}}(t_\mu)(\hat{\boldsymbol{J}}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\right]\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\mathcal{F}[1; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\right.$$
$$\left. -\frac{1}{2}\eta(t_\mu)\boldsymbol{J}(t_\mu)\mathcal{F}^2[1; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\right\} + \mathcal{O}(N^{-2})$$

The evolution of the observable $\omega[\boldsymbol{J}] = \hat{\boldsymbol{J}}\cdot\boldsymbol{B}$ is thus given by

$$\omega[\boldsymbol{J}(t_\mu + \frac{1}{N})] = \omega[\boldsymbol{J}(t_\mu)] + \frac{1}{N}\eta(t_\mu)\left\{\left[|\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu| - \omega[\boldsymbol{J}(t_\mu)](\hat{\boldsymbol{J}}(t_\mu)\cdot\boldsymbol{\xi}^\mu)\ \mathrm{sgn}(\boldsymbol{B}\cdot\boldsymbol{\xi}^\mu)\right]\mathcal{F}[1; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\right.$$
$$\left. -\frac{1}{2}\omega\eta(t_\mu)\mathcal{F}^2[1; \boldsymbol{J}(t_\mu)\cdot\boldsymbol{\xi}^\mu, \boldsymbol{B}\cdot\boldsymbol{\xi}^\mu]\right\} + \mathcal{O}(N^{-2})$$

Following the procedure of section 1.2 then leads to

$$\frac{d}{dt}\omega = \eta\langle [|y| - \omega x\ \mathrm{sgn}(y)]\,\mathcal{F}[1; x, y] - \frac{1}{2}\omega\eta^2\langle \mathcal{F}^2[1; x, y]\rangle \tag{4.39}$$

which is identical to equation (4.37).

Finally we convert equation (4.37) into a dynamical equation for the error E, using (4.15), which gives the final result

$$\frac{d}{dt}E = -\frac{\eta}{\pi\sin(\pi E)}\langle [|y| - \cos(\pi E)x\ \mathrm{sgn}(y)]\,\mathcal{F}[1; x, y]\rangle + \frac{\eta^2}{2\pi\tan(\pi E)}\langle \mathcal{F}^2[1; x, y]\rangle \tag{4.40}$$

with averages defined with respect to the distribution (4.14), in which $\omega = \cos(\pi E)$.

For spherical models described by either of the equivalent classes of on-line rules (4.36) or (4.38) the evolution of the error is described by a single first-order non-linear differential equation, rather than a pair of coupled non-linear differential equations. This will allow us to push the analysis further, but the price we pay is that of a loss in generality.

Optimal Time-Dependent Learning Rates. We wish to optimise the approach to the $E = 0$ state of our macroscopic equations, by choosing a suitable time-dependent learning rate. Let us distinguish between the possible situations we can find ourselves in. If our learning rule is of the general form (4.6), without spherical normalisation, we have two coupled macroscopic equations:

$$\frac{d}{dt}J = \eta\langle x \,\mathrm{sgn}(y)\mathcal{F}[J; Jx, y]\rangle + \frac{\eta^2}{2J}\langle \mathcal{F}^2[J; Jx, y]\rangle \tag{4.41}$$

$$\frac{d}{dt}E = -\frac{\eta}{J\pi\sin(\pi E)}\langle [|y| - \cos(\pi E)x \,\mathrm{sgn}(y)]\,\mathcal{F}[J; Jx, y]\rangle + \frac{\eta^2}{2\pi J^2\tan(\pi E)}\langle \mathcal{F}^2[J; Jx, y]\rangle \tag{4.42}$$

which are obtained by combining (4.12,4.13) with (4.15). The probability distribution (4.14) with which the averages are computed depends on E only, not on J. If, on the other hand, we complement the rule (4.6) with weight vector normalisation as in (4.36) or (4.38) (the spherical rules), we obtain a single equation for E only:

$$\frac{d}{dt}E = -\frac{\eta}{\pi\sin(\pi E)}\langle [|y| - \cos(\pi E)x \,\mathrm{sgn}(y)]\,\mathcal{F}[1; x, y]\rangle + \frac{\eta^2}{2\pi\tan(\pi E)}\langle \mathcal{F}^2[1; x, y]\rangle \tag{4.43}$$

Since equation (4.43) is autonomous (there are no dynamical variables other than E), the optimal choice of the function $\tilde{\eta}(t)$ (i.e. the one that generates the fastest decay of the error E) is obtained by simply minimising the temporal derivative of the error *at each time-step*:

$$\forall t \geq 0: \qquad \frac{\partial}{\partial\tilde{\eta}(t)}\left[\frac{d}{dt}E\right] = 0 \tag{4.44}$$

which is called the 'greedy' recipe. Note, however, that the same is true for equation (4.42) if we restrict ourselves to rules with the property that $\mathcal{F}[J; Jx, y] = \gamma(J)\mathcal{F}[1; x, y]$ for some function $\gamma(J)$, such as the Hebbian ($\gamma(J) = 1$), perceptron ($\gamma(J) = 1$) and AdaTron ($\gamma(J) = J$) rules. This property can also be written as

$$\frac{\partial}{\partial x}\frac{\mathcal{F}[J; Jx, y]}{\mathcal{F}[1; x, y]} = \frac{\partial}{\partial y}\frac{\mathcal{F}[J; Jx, y]}{\mathcal{F}[1; x, y]} = 0 \tag{4.45}$$

For rules which obey (4.45) we can simply write the time-dependent learning rate as $\eta = \tilde{\eta}J/\gamma(J)$, such that equations (4.41,4.42) acquire the form:

$$\frac{d}{dt}\log J = \tilde{\eta}\langle x \,\mathrm{sgn}(y)\mathcal{F}[1; x, y]\rangle + \frac{1}{2}\tilde{\eta}^2\langle \mathcal{F}^2[1; x, y]\rangle \tag{4.46}$$

$$\frac{d}{dt}E = -\frac{\tilde{\eta}}{\pi\sin(\pi E)}\langle [|y| - \cos(\pi E)x \,\mathrm{sgn}(y)]\,\mathcal{F}[1; x, y]\rangle + \frac{\tilde{\eta}^2}{2\pi\tan(\pi E)}\langle \mathcal{F}^2[1; x, y]\rangle \tag{4.47}$$

In these cases, precisely since we are free to choose the function $\tilde{\eta}(t)$ as we wish, the evolution of J decouples from our problem of optimising the evolution of E. For learning rules where $\mathcal{F}[J;Jx,y]$ truly depends on J, on the other hand (i.e. where (4.45) does not hold), optimisation of the error relaxation is considerably more difficult, and is likely to depend on the particular time t for which one wants to minimise $E(t)$. We will not deal with such cases here.

If the 'greedy' recipe applies (for spherical rules and for ordinary ones with the property (4.45)) working out the derivative in (4.44) immediately gives us

$$\tilde{\eta}(t)_{\rm opt}=\frac{\langle\{|y|-\cos(\pi E)x\ {\rm sgn}(y)\}\mathcal{F}[1;x,y]\rangle}{\cos(\pi E)\langle\mathcal{F}^2[1;x,y]\rangle} \tag{4.48}$$

Insertion of this choice into equation (4.40) subsequently leads to

$$\left.\frac{dE}{dt}\right|_{\rm opt}=-\frac{\langle\{|y|-\cos(\pi E)x\ {\rm sgn}(y)\}\mathcal{F}[1;x,y]\rangle^2}{2\pi\sin(\pi E)\cos(\pi E)\langle\mathcal{F}^2[1;x,y]\rangle} \tag{4.49}$$

These and subsequent expressions we will write in terms of $\tilde{\eta}$, defined as $\tilde{\eta}(t)=\eta(t)$ for the spherical learning rules and as $\tilde{\eta}(t)=\eta(t)J(t)/\gamma(J(t))$ for the non-spherical learning rules. We will now work out the details of the results (4.48,4.49) upon making the familiar choices for the function $\mathcal{F}[\ldots]$: the Hebbian, perceptron and AdaTron rules.

For the (ordinary and spherical) Hebbian rules, corresponding to $\mathcal{F}[J;Jx,y]=1$, the various Gaussian integrals in (4.48,4.49) are the same as those we already did (analytically) in the case of constant learning rate η. Substitution of the outcomes of the integrals (see appendix) into the equations (4.48,4.49) gives

$$\tilde{\eta}_{\rm opt}=\sqrt{\frac{2}{\pi}}\,\frac{\sin^2(\pi E)}{\cos(\pi E)}\qquad\qquad \left.\frac{dE}{dt}\right|_{\rm opt}=-\frac{\sin^3(\pi E)}{\pi^2\cos(\pi E)}$$

The equation for the error E can be solved explicitly, giving (to be verified by substitution):

$$t(E)=\frac{1}{2}\pi\sin^{-2}(\pi E)-\frac{1}{2}\pi\sin^{-2}(\pi E_0) \tag{4.50}$$

The asymptotic behaviour of the process follows from expansion of (4.50) for small E, and gives

$$E_{\rm opt}\sim\frac{1}{\sqrt{2\pi t}}\qquad\qquad \tilde{\eta}_{\rm opt}\sim\sqrt{\frac{\pi}{2}}\,\frac{1}{t}\qquad\qquad (t\to\infty)$$

Asymptotically there is nothing to be gained by choosing the optimal time-dependent learning rate, since the same asymptotic form for E was also obtained for constant η (see (4.22)). Note that the property $\mathcal{F}[J;Jx,y]=\mathcal{F}[1;x,y]$ of the Hebbian recipe guarantees that the result (4.50) applies to both the ordinary and the spherical Hebbian rule. The only difference between the two cases is in the definition of $\tilde{\eta}$: for the ordinary (non-spherical) version $\tilde{\eta}(t)=\eta(t)/J(t)$, whereas for the spherical version $\tilde{\eta}(t)=\eta(t)$.

We move on to the (ordinary and spherical) perceptron learning rules, where $\mathcal{F}[J;Jx,y]=\theta[-xy]$, with time-dependent learning rates $\eta(t)$ which we aim to optimise. As in the Hebbian case all integrals occurring in (4.48,4.49) upon substitution of the present choice $\mathcal{F}[J;Jx,y]=\theta[-xy]$ have been done already (see the appendix) . Insertion of the outcomes of these integrals into (4.48,4.49) gives

$$\tilde{\eta}_{\rm opt}=\frac{\sin^2(\pi E)}{\sqrt{2\pi}E\cos(\pi E)}\qquad\qquad \left.\frac{dE}{dt}\right|_{\rm opt}=-\frac{\sin^3(\pi E)}{4\pi^2E\cos(\pi E)}$$

Again the non-linear differential equation describing the evolution of the error E can be solved exactly:

$$t(E) = \frac{2[\pi E + \sin(\pi E)\cos(\pi E)]}{\sin^2(\pi E)} - \frac{2[\pi E_0 + \sin(\pi E_0)\cos(\pi E_0)]}{\sin^2(\pi E_0)} \tag{4.51}$$

Expansion of (4.51) for small E gives the asymptotic behaviour

$$E_{\text{opt}} \sim \frac{4}{\pi t} \qquad\qquad \tilde{\eta}_{\text{opt}} \sim \frac{2\sqrt{2\pi}}{t} \qquad\qquad (t \to \infty)$$

which is identical to that found in the beginning of this section, i.e. equations (4.31,4.35), upon exploring the consequences of making two simple ad-hoc choices for the time-dependent learning rate (since $\tilde{\eta} = \eta/J$). As with the Hebbian rule the property $\mathcal{F}[J; Jx, y] = \mathcal{F}[1; x, y]$ of the perceptron recipe guarantees that the result (4.51) applies to both the ordinary and the spherical version.

Finally we try to optimise the learning rate for the spherical AdaTron learning rule, corresponding to the choice $\mathcal{F}[J; Jx, y] = |Jx|\theta[-xy]$. Working out the averages in (4.48,4.49) again does not require doing any new integrals. Using those already encountered in analysing the AdaTron rule with constant learning rate (to be found in the appendix), we obtain

$$\tilde{\eta}_{\text{opt}} = \frac{\sin^3(\pi E)}{\pi}\left[E\cos(\pi E) - \frac{\cos^2(\pi E)\sin(\pi E)}{\pi}\right]^{-1}$$

$$\left.\frac{dE}{dt}\right|_{\text{opt}} = -\frac{\sin^5(\pi E)}{2\pi^2\cos(\pi E)}\left[\frac{1}{\pi E - \cos(\pi E)\sin(\pi E)}\right]$$

(note that in both versions, ordinary and spherical, of the AdaTron rule we simply have $\tilde{\eta}(t) = \eta(t)$). It will no longer come as a surprise that also this equation for the evolution of the error allows for analytical solution:

$$t(E) = \frac{\pi}{8}\left[\frac{4\pi E - \sin(4\pi E)}{\sin^4(\pi E)} - \frac{4\pi E_0 - \sin(4\pi E_0)}{\sin^4(\pi E_0)}\right] \tag{4.52}$$

Asymptotically we find, upon expanding (4.52) for small E, a relaxation of the form

$$E_{\text{opt}} \sim \frac{4}{3t} \qquad\qquad \tilde{\eta}_{\text{opt}} \sim \frac{3}{2} \qquad\qquad (t \to \infty)$$

So for the AdaTron rule the asymptotic behaviour for optimal time-dependent learning rate η is identical to that found for optimal *constant* learning rate η (which is indeed $\eta = \frac{3}{2}$, see (4.30)). As with the previous two rules, the property $\mathcal{F}[J; Jx, y] = J\mathcal{F}[1; x, y]$ of the AdaTron recipe guarantees that the result (4.50) applies to both the ordinary and the spherical version.

It is quite remarkable that the simple perceptron learning rule, which came out at the bottom of the league among the three learning rules considered so far in the case of having constant learning rates, all of a sudden comes out with 'douze points' as soon as we allow for optimised time-dependent learning rates. It is in addition quite satisfactory that in a number of cases one can actually find an explicit expression for the relation $t(E)$ between the duration of the learning stage and the generalization error achieved, i.e. equations (4.34,4.50,4.51,4.52).

Optimal On-line Learning Rules. We need not restrict our optimisation attempts to varying the learning rate η only, but we can also vary the full form $\eta\mathcal{F}[J;Jx,y]$ of the learning rule. The aim, as always, is to minimise the generalisation error, but there will be limits to what is achievable. So far all examples of on-line learning rules we have studied gave an asymptotic relaxation of the error of the form $E \sim t^{-q}$ with $q \leq 1$. It can be shown using general probabilistic arguments that

$$\lim_{t\to\infty} tE(t) \geq 0.44\ldots \tag{4.53}$$

No on-line learning rule can violate (4.53)[7]. On the other hand: we have already encountered several rules with at least the optimal power $E \sim t^{-1}$. The optimal on-line learning rule is thus one which gives asymptotically $E \sim A/t$, but with the smallest value of A possible.

The function $\mathcal{F}[J;Jx,y]$ in the learning rules is allowed to depend only on the *sign* of the teacher field $y = \boldsymbol{B}\cdot\boldsymbol{\xi}$, not on its magnitude, since otherwise it would describe a situation where considerably more than just the answers $T(\boldsymbol{\xi}) = \text{sgn}[\boldsymbol{B}\cdot\boldsymbol{\xi}]$ of the teacher are used for updating the parameters of the student. One can easily see that using unavailable information indeed violates (4.53). Suppose, for instance, we would consider spherical on-line rules, i.e. (4.36) or (4.38), and make the forbidden choice

$$\eta\mathcal{F}[1;x,y] = \frac{|y| - \cos(\pi E)x\ \text{sgn}(y)}{\cos(\pi E)}$$

We would then find for the corresponding equation (4.43) describing the evolution of the error E for $N \to \infty$:

$$\frac{d}{dt}E = -\frac{\langle[|y| - \cos(\pi E)x\ \text{sgn}(y)]^2\rangle}{2\pi\sin(\pi E)\cos(\pi E)}$$

(with averages as always calculated with the distribution (4.14)) from which it follows, upon using the Gaussian intregrals done in the appendix:

$$\frac{d}{dt}E = -\frac{\tan(\pi E)}{2\pi}$$

This produces exponential decay of the error, and thus indeed violates (4.53).

Taking into account the restrictions on available information, and anticipating the form subsequent expressions will take, we write the function $\mathcal{F}[J;Jx,y]$ (which we will be varying, and which we will also allow to have an explicit time-dependence[8]) in the following form

$$\eta\mathcal{F}[J;Jx,y] = \begin{cases} J\mathcal{F}_+(x,t) & \text{if } y > 0 \\ J\mathcal{F}_-(x,t) & \text{if } y < 0 \end{cases} \tag{4.54}$$

[7] This will be different for graded-response perceptrons.

[8] By allowing for an explit time-dependence, we can drop the dependence on J in $\mathcal{F}[J;Jx,y]$ if we wish, without loss of generality, since J is itself just some function of time.

If our learning rule is of the general form (4.6), without spherical normalisation, the coupled equations (4.41,4.42) describe the macroscopic dynamics. For the spherical rules (4.36,4.38) we have the single macroscopic equation (4.43). Both (4.42) and (4.43) now acquire the form

$$\frac{d}{dt}E = -\frac{1}{\pi\sin(\pi E)}\Big\{\langle(y-\omega x)\theta[y]\mathcal{F}_{+}(x,t)\rangle - \langle(y-\omega x)\theta[-y]\mathcal{F}_{-}(x,t)\rangle$$
$$-\frac{1}{2}\omega\langle\theta[y]\mathcal{F}_{+}^{2}(x,t)\rangle - \frac{1}{2}\omega\langle\theta[-y]\mathcal{F}_{-}^{2}(x,t)\rangle\Big\} \qquad (4.55)$$

with the usual short-hand $\omega = \cos(\pi E)$ and with averages calculated with the (time-dependent) distribution (4.14). To simplify notation we now introduce the two functions

$$\int dy\ \theta[y]P(x,y) = \Omega(x,t) \qquad\qquad \int dy\ \theta[y](y-\omega x)P(x,y) = \Delta(x,t)$$

and hence, using the symmetry $P_t(x,y) = P_t(-x,-y)$, equation (4.55) acquires the compact form

$$\frac{d}{dt}E = -\frac{1}{\pi\sin(\pi E)}\int dx\,\Big\{\Delta(x,t)\mathcal{F}_{+}(x,t) - \frac{1}{2}\omega\Omega(x,t)\mathcal{F}_{+}^{2}(x,t)\Big\}$$
$$-\frac{1}{\pi\sin(\pi E)}\int dx\,\Big\{\Delta(-x,t)\mathcal{F}_{-}(x,t)\rangle - \frac{1}{2}\omega\Omega(-x,t)\mathcal{F}_{-}^{2}(x,t)\Big\} \qquad (4.56)$$

Since there is only one dynamical variable, the error E, our optimisation problem is solved by the 'greedy' recipe which here involves functional derivatives:

$$\forall x,\ \forall t: \qquad \frac{\delta}{\delta\mathcal{F}_{+}(x,t)}\left[\frac{dE}{dt}\right] = \frac{\delta}{\delta\mathcal{F}_{-}(x,t)}\left[\frac{dE}{dt}\right] = 0$$

with the solution

$$\mathcal{F}_{+}(x,t) = \frac{\Delta(x,t)}{\omega\Omega(x,t)} \qquad\qquad \mathcal{F}_{-}(x,t) = \frac{\Delta(-x,t)}{\omega\Omega(-x,t)} = \mathcal{F}_{+}(-x,t)$$

Substitution of this solution into (4.56) gives the corresponding law describing the optimal error evolution of (ordinary and spherical) on-line rules:

$$\left.\frac{dE}{dt}\right|_{\rm opt} = -\frac{1}{\pi\sin(\pi E)\cos(\pi E)}\int dx\frac{\Delta^{2}(x,t)}{\Omega(x,t)}$$

Explicit calculation of the integrals $\Delta(x,t)$ and $\Omega(x,t)$ (see appendix) gives:

$$\Delta(x,t) = \frac{\sin(\pi E)}{2\pi}e^{-\frac{1}{2}x^{2}/\sin^{2}(\pi E)} \qquad\qquad \Omega(x,t) = \frac{e^{-\frac{1}{2}x^{2}}}{2\sqrt{2\pi}}\Big[1+{\rm erf}\left(x/\sqrt{2}\tan(\pi E)\right)\Big]$$

with which we finally obtain an explicit expression for the optimal form of the learning rule, via (4.54), as well as for the dynamical law describing the corresponding error evolution:

$$\eta\mathcal{F}[J;Jx,y]_{\rm opt} = \sqrt{\frac{2}{\pi}}\ \frac{J\tan(\pi E)e^{-\frac{1}{2}x^{2}/\tan^{2}(\pi E)}}{1+\ {\rm sgn}(xy){\rm erf}\left(|x|/\sqrt{2}\tan(\pi E)\right)} \qquad (4.57)$$

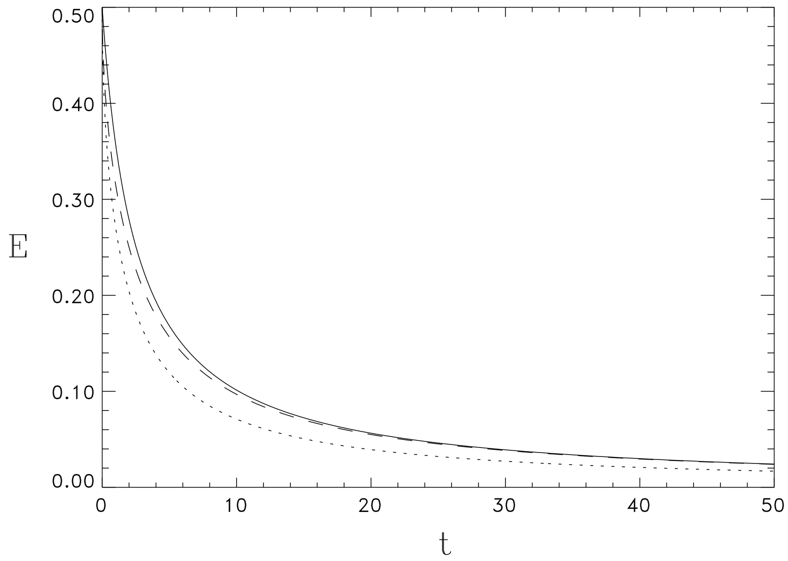


Figure 4.9: Evolution of the error E for three on-line learning rules: Perceptron rule with a learning rate such that $J(t) = 1$ for all $t \geq 0$ (solid line), Perceptron rule with optimal learning rate (dashed line) and the optimal spherical learning rule (dotted line). Initial state: $E(0) = \frac{1}{2}$ and $J(0) = 1$. The curves for the Perceptron rules are given by (4.34) and (4.51). The curve for the optimal spherical rule was obtained by numerical solution of equation ((4.58).

$$\left.\frac{dE}{dt}\right|_{\text{opt}} = -\frac{\tan^2(\pi E)}{\pi^2\sqrt{2\pi}} \int dx \; \frac{e^{-\frac{1}{2}x^2[1+\cos^2(\pi E)]/\cos^2(\pi E)}}{1 + \text{erf}(x/\sqrt{2})} \tag{4.58}$$

The asymptotic form of the error relaxation towards the $E = 0$ state follows from expansion of equation (4.58) for small E, which gives

$$\frac{dE}{dt} = -E^2 \int dy \frac{e^{-y^2}}{\sqrt{2\pi}[1 + \text{erf}(y/\sqrt{2})]} + \mathcal{O}(E^4)$$

so that we can conclude that the optimum asymptotic decay for on-line learning rules (whether spherical or non-spherical) is given by $E \sim \frac{A}{t}$ for $t \to \infty$, with

$$A^{-1} = \int dx \; \frac{e^{-x^2}}{\sqrt{2\pi}[1 + \text{erf}(x/\sqrt{2})]}$$

Numerical evaluation of this integral (which is somewhat delicate due to the behaviour of the integrand for $y \to -\infty$) finally gives

$$E \sim \frac{0.883\ldots}{t} \qquad\qquad (t \to \infty)$$

It is instructive to investigate briefly the form of the optimal learning rule (4.57) for large values of E (as in the initial stages of learning processes) and for small values of E (as in the

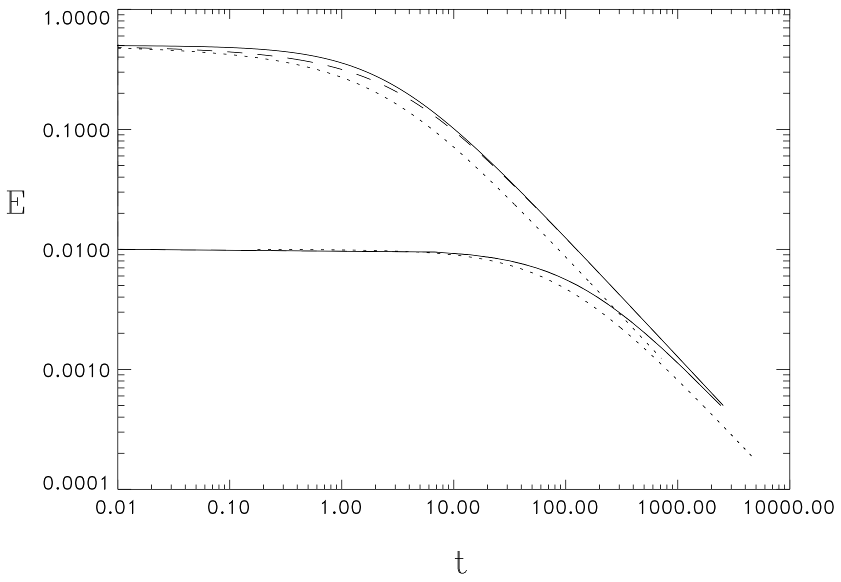


Figure 4.10: Evolution of the error E for the on-line Perceptron rule with a learning rate such that $J(t) = 1$ for all $t \geq 0$ (solid line), the on-line Perceptron rule with optimal learning rate (dashed line) and the optimal spherical on-line learning rule (dotted line). Initial states: $(J, E) = (1, \frac{1}{2})$ (upper curves), and $(J, E) = (1, \frac{1}{100})$ (lower curves). The curves for the Perceptron rules are given by (4.34) and (4.51). The curves for the optimal spherical rule were obtained by numerical solution of equation (4.58).

final stages of learning processes). Initially we find

$$\lim_{E \uparrow \frac{1}{2}} \frac{\eta \mathcal{F}[J; Jx, y]_{\text{opt}}}{\tan(\pi E)} = J\sqrt{\frac{2}{\pi}}$$

which describes a Hebbian-type learning rule with diverging learning rate (note that $\tan(\pi E) \to \infty$ for $E \uparrow \frac{1}{2}$). In contrast, in the final stages the optimal learning rule (4.57) acquires the form

$$\lim_{E \downarrow 0} \eta \mathcal{F}[J; Jx, y]_{\text{opt}} = \frac{J|x|}{\sqrt{\pi}}\, \theta[-xy] \lim_{z \to \infty} \frac{e^{-z^2}}{z\,[1 - \text{erf}(z)]} = J|x|\theta[-xy]$$

which is the AdaTron learning rule with learning rate $\eta = 1$[9].

In figures 4.9 (short times and ordinary axes) and 4.10 (large times and log-log axes) we finally compare the evolution of the error for the optimal on-line learning rule (4.57) with the two on-line learning rules which so far were found to give the fastest relaxation: the perceptron rule with normalising time-dependent learning rate (giving the error of (4.34)),

[9] The reason that, in spite of the asymptotic equivalence of the two rules, the optimal rule does not asymptotically give the same relaxation of the error E as the AdaTron rule is that in order to determine the asymptotics one has to take the limit $E \to 0$ in the full macroscopic differential equation for E, which, in addition to the function $\mathcal{F}[\ldots]$ defining the learning rule, involves the Gaussian probability distribution (4.14) which depends on E in a non-trivial way, especially near $E = 0$.

and the perceptron rule with optimal time-dependent learning rate (giving the error of (4.51)). This in order to assess whether choosing the optimal on-line learning rule (4.57) rather than its simpler competitors is actually worth the effort. The curves for the optimal on-line rule were obtained by numerical solution of equation (4.58).

Summary in a Table. We close this section with an overview of some of the results on on-line learning in perceptrons described/derived so far. The upper part of this table contains results for specific learning rules with either arbitrary constant learning rates η (first column), optimal constant learning rate η (second column), and where possible, a time-dependent learning rate $\eta(t)$ chosen such as to realise the normalisation $J(t) = 1$ for all t. The lower part of the table gives results for specific learning rules with optimised time dependent learning rates $\eta(t)$, as well as lower bounds on the asymptotic generalization error.

GENERALIZATION ERROR IN PERCEPTRONS WITH ON-LINE LEARNING RULES

	Constant learning rate η		Variable η
Rule	Asymptotic decay for constant η	Optimal asymptotic decay for constant η	η chosen to normalise J
Hebbian	$E \sim \frac{1}{\sqrt{2\pi}} t^{-1/2}$ for $\eta > 0$	$E \sim \frac{1}{\sqrt{2\pi}} t^{-1/2}$ for $\eta > 0$	N/A
Perceptron	$E \sim (\frac{2}{3})^{1/3} \pi^{-1} t^{-1/3}$ for $\eta > 0$	$E \sim (\frac{2}{3})^{1/3} \pi^{-1} t^{-1/3}$ for $\eta > 0$	$E \sim \frac{4}{\pi} t^{-1}$
AdaTron	$E \sim (\frac{3}{3\eta - \eta^2}) t^{-1}$ for $0 < \eta < 3$	$E \sim \frac{4}{3} t^{-1}$ for $\eta = \frac{3}{2}$	$E \sim \frac{3}{2} t^{-1}$

OPTIMAL GENERALIZATION

	Optimal time-dependent learning rate η	
Rule	Generalization error for optimal time-dependent η	Asymptotics
Hebbian	$t = \frac{\pi}{2}[\frac{1}{\sin^2(\pi E)} - \frac{1}{\sin^2(\pi E_0)}]$	$E \sim \frac{1}{\sqrt{2\pi}} t^{-1/2}$
Perceptron	$t = 2[\frac{\pi E + \sin(\pi E)\cos(\pi E)}{\sin^2(\pi E)} - \frac{\pi E_0 + \sin(\pi E_0)\cos(\pi E_0)}{\sin^2(\pi E_0)}]$	$E \sim \frac{4}{\pi} t^{-1}$
AdaTron	$t = \frac{\pi}{8}[\frac{4\pi E - \sin(4\pi E)}{\sin^4(\pi E)} - \frac{4\pi E_0 - \sin(4\pi E_0)}{\sin^4(\pi E_0)}]$	$E \sim \frac{4}{3} t^{-1}$
Lower bound for on-line learning (asymptotics of the optimal learning rule)		$E \sim 0.88 t^{-1}$
Lower bound for any learning rule		$E \sim 0.44 t^{-1}$

Appendix A

The δ-Distribution

Definition. There are several ways of introducing the δ-distribution. Here we will go for an intuitive definition first, and a formal one later. We define the δ-distribution as the probability distribution $\delta(x)$ corresponding to a random variable in the limit where the randomness in the variable vanishes. If x is 'distributed' around zero, this implies

$$\int dx\; f(x)\delta(x) = f(0) \qquad \text{for any function } f$$

The problem arises when we want to actually write down an expression for $\delta(x)$. Intuitively one could think of writing something like

$$\delta(x) = \lim_{\Delta\to 0} G_\Delta(x) \qquad G_\Delta(x) = \frac{1}{\Delta\sqrt{2\pi}} e^{-\frac{1}{2}x^2/\Delta^2} \tag{A.1}$$

This is not a true function in a mathematical sense; $\delta(x)$ is zero for $x \neq 0$ and $\delta(0) = \infty$. The way to interpret and use expressions like (A.1) is to realise that $\delta(x)$ only has a meaning when appearing inside an integration. One then takes the limit $\Delta \to 0$ *after* performing the integration. Upon adopting this convention, we can use (A.1) to derive the following properties (for sufficiently well-behaved and differentiable functions f[1]):

$$\int dx\; \delta(x)f(x) = \lim_{\Delta\to 0}\int dx\; G_\Delta(x)f(x) = \lim_{\Delta\to 0}\int \frac{dx}{\sqrt{2\pi}}\; e^{-\frac{1}{2}x^2} f(\Delta x) = f(0)$$

$$\int dx\; \delta'(x)f(x) = \lim_{\Delta\to 0}\int dx \left\{\frac{d}{dx}\left[G_\Delta(x)f(x)\right] - G_\Delta(x)f'(x)\right\}$$

$$= \lim_{\Delta\to 0}\left[G_\Delta(x)f(x)\right]_{-\infty}^{\infty} - f'(0) = -f'(0)$$

both can be summarised in and generalised to the single expression:

$$\int dx\; f(x)\frac{d^n}{dx^n}\delta(x) = (-1)^n \lim_{x\to 0}\frac{d^n}{dx^n}f(x) \qquad (n = 0, 1, 2, \ldots) \tag{A.2}$$

[1] The conditions on the so-called 'test-functions' f can be properly formalised; this being not a course on distribution theory, here we just concentrate on the basic ideas and properties

Equivalently we can take the result (A.2) as our definition of the δ-distribution.

$\delta(x)$ as Solution of the Liouville Equation. Here we prove that the δ-distribution can be used to represent the solution of the so-called Liouville equation:

$$\frac{\partial}{\partial t}P_t(x) = -\frac{\partial}{\partial x}\left[P_t(x)F(x)\right] \tag{A.3}$$

The general solution of (A.3) is

$$P_t(x) = \int dx_0\ P_0(x_0)\delta[x - x^\star(t;x_0)] \tag{A.4}$$

in which $x^\star(t;x_0)$ is the solution of the ordinary differential equation

$$\frac{d}{dt}x^\star(t) = F(x^\star(t)) \qquad x^\star(0) = x_0 \tag{A.5}$$

In particular, if $P_0(x_0)$ is a δ-distribution in x_0, the general solution will remain a δ-distribution in x for all times: $P_t(x) = \delta[x{-}x^\star(t;x_0)]$. The proof that (A.4) is true consists of showing that both sides of (A.3) give the same result inside integrals, if we insert the proposed solution (A.4):

$$\begin{aligned}
&\int dx\ f(x)\left\{\frac{\partial}{\partial t}P_t(x) + \frac{\partial}{\partial x}\left[P_t(x)F(x)\right]\right\}\\
&\qquad = \frac{\partial}{\partial t}\int dx\ f(x)P_t(x) + \left[f(x)P_t(x)F(x)\right]_{-\infty}^{\infty} - \int dx\ P_t(x)F(x)f'(x)\\
&\qquad = \int dx_0\ P_0(x_0)\left\{\frac{\partial}{\partial t}f(x^\star(t;x_0)) - F(x^\star(t;x_0))f'(x^\star(t;x_0))\right\}\\
&\qquad = \int dx_0\ P_0(x_0)f'(x^\star(t;x_0))\left\{\frac{d}{dt}x^\star(t;x_0) - F(x^\star(t;x_0))\right\} = 0
\end{aligned}$$

Representations, Relations, Generalisations. We can use the definitions of Fourier transforms and inverse Fourier transforms to obtain an integral representation of the δ-distribution:

$$\mathcal{F}:\ f(x) \to \hat{f}(k) \qquad \hat{f}(k) = \int dx\ e^{-2\pi ikx}f(x)$$

$$\mathcal{F}^{-1}:\ \hat{f}(k) \to f(x) \qquad f(x) = \int dk\ e^{2\pi ikx}\hat{f}(k)$$

In combination these relations give the identity:

$$f(x) = \int dk\ e^{2\pi ikx}\int dy\ e^{-2\pi iky}f(y)$$

Application to $f(x) = \delta(x)$ gives:

$$\delta(x) = \int dk\ e^{2\pi ikx} = \int \frac{dk}{2\pi}\ e^{ikx} \tag{A.6}$$

Another useful relation is the following one, which relates the δ-distribution to the step-function:

$$\delta(x) = \frac{d}{dx}\theta(x) \tag{A.7}$$

This we prove by showing that both have the same effect inside an integration (with an arbitrary test-function):

$$\int dx \left[\delta(x) - \frac{d}{dx}\theta(x)\right] f(x) = f(0) - \lim_{\epsilon\to 0}\int_{-\epsilon}^{\epsilon} dx \left\{\frac{d}{dx}\left[\theta(x)f(x)\right] - f'(x)\theta(x)\right\}$$

$$= f(0) - \lim_{\epsilon\to 0}\left[f(\epsilon) - 0\right] + \lim_{\epsilon\to 0}\int_{0}^{\epsilon} dx\ f'(x) = 0$$

Finally, the following generalisation is straightforward:

$$\boldsymbol{x} \in \mathcal{R}^N: \qquad \delta(\boldsymbol{x}) = \prod_{i=1}^{N} \delta(x_i) \tag{A.8}$$

Appendix B

Gaussian Integrals

In this appendix we give brief derivations of those integrals used in the cahpter on learning dynamics that turn out to be easy, and give the appropriate reference for finding the nasty ones. All involve the following Gaussian distribution:

$$\langle f(x,y)\rangle = \int dxdy\ f(x,y)P(x,y) \qquad\qquad P(x,y) = \frac{1}{2\pi\sqrt{1-\omega^2}}e^{-\frac{1}{2}[x^2+y^2-2xy\omega]/(1-\omega^2)}$$

I: $I_1 = \langle|y|\rangle$

$$I_1 = \int \frac{dy}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}|y| = \sqrt{\frac{2}{\pi}}$$

II: $I_2 = \langle x\ \mathrm{sgn}(y)\rangle$

$$I_2 = -\int \frac{dxdy}{2\pi\sqrt{1-\omega^2}}\ \mathrm{sgn}(y)e^{-\frac{1}{2}[y^2-2\omega xy]/(1-\omega^2)}(1-\omega^2)\frac{\partial}{\partial x}e^{-\frac{1}{2}x^2/(1-\omega^2)}$$

$$= \int \frac{dxdy}{2\pi\sqrt{1-\omega^2}}e^{-\frac{1}{2}[x^2+y^2-2\omega xy]/(1-\omega^2)}\ \mathrm{sgn}(y)\omega y$$

$$= \omega\langle|y|\rangle = \omega\sqrt{\frac{2}{\pi}}$$

III: $I_3 = \langle\theta[-xy]\rangle$

$$I_3 = \int_0^\infty\!\!\int_0^\infty \frac{dxdy}{\pi\sqrt{1-\omega^2}}e^{-\frac{1}{2}[x^2+y^2+2\omega xy]/(1-\omega^2)} = \frac{\sqrt{1-\omega^2}}{\pi}\int_0^\infty\!\!\int_0^\infty dxdy\, e^{-\frac{1}{2}[x^2+y^2+2\omega xy]}$$

Introduce polar coordinates $(x,y) = r(\cos\phi, \sin\phi)$:

$$I_3 = \frac{\sqrt{1-\omega^2}}{\pi}\int_0^{\pi/2} d\phi \int_0^\infty dr\, re^{-\frac{1}{2}r^2[1+\omega\sin(2\phi)]}$$

$$= \frac{\sqrt{1-\omega^2}}{2\pi}\int_0^\pi \frac{d\phi}{1+\omega\sin(\phi)} = \frac{1}{\pi}\left[\frac{\pi}{2} - \arctan\left(\frac{\omega}{\sqrt{1-\omega^2}}\right)\right]$$

(the last integral can be found in I.S. Gradshteyn and I.M. Ryzhik,'Table of Integrals, Series amd Products', 1980, Academic Press). Finally, using $\cos[\frac{\pi}{2}-\psi]=\sin\psi$, we find

$$I_3 = \frac{1}{\pi}\arccos(\omega)$$

IV: $I_4 = \langle x\ \ \mathrm{sgn}(y)\,\theta[-xy]\rangle$

$$\begin{aligned} I_4 &= -\frac{1-\omega^2}{\pi}\int_0^\infty dx\,x\,e^{-\frac{1}{2}x^2}\int_0^\infty dy\,e^{-\frac{1}{2}[y+\omega x]^2+\frac{1}{2}\omega^2x^2} \\ &= \frac{1}{\pi}\left[e^{-\frac{1}{2}x^2}\int_{\omega x/\sqrt{1-\omega^2}}^\infty dy\,e^{-\frac{1}{2}y^2}\right]_0^\infty - \frac{1}{\pi}\int_0^\infty dx\,e^{-\frac{1}{2}y^2}\frac{\partial}{\partial x}\int_{\omega x/\sqrt{1-\omega^2}}^\infty dy\,e^{-\frac{1}{2}y^2} \\ &= -\frac{1}{\sqrt{2\pi}} + \frac{\omega}{\pi\sqrt{1-\omega^2}}\int_0^\infty dx\,e^{-\frac{1}{2}x^2/(1-\omega^2)} = \frac{\omega-1}{\sqrt{2\pi}} \end{aligned}$$

V: $I_5 = \langle |y|\theta[-xy]\rangle$

$$I_5 = \int_0^\infty\!\!\int_0^\infty dxdy\ y[P(x,-y)+P(-x,y)] = \frac{1-\omega}{\sqrt{2\pi}}$$

VI: $I_6 = \langle x^2\theta[-xy]\rangle$

$$\begin{aligned} I_6 &= \frac{1}{\pi\sqrt{1-\omega^2}}\int_0^\infty\int_0^\infty dxdy\,x^2\,e^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)} \\ &= \frac{1}{2\pi\sqrt{1-\omega^2}}\int_0^\infty\int_0^\infty dxdy\,(x^2+y^2)\,e^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)} \end{aligned}$$

We switch to polar coordinates $(x,y) = r(\cos\theta,\sin\theta)$, and subsequently substitute $t = r^2[1+\omega\sin(2\theta)]/[1-\omega^2]$:

$$\begin{aligned} I_6 &= \frac{1}{2\pi\sqrt{1-\omega^2}}\int_0^{\pi/2} d\theta\int_0^\infty dr\,r^3\,e^{-\frac{1}{2}[r^2+2\omega r^2\cos\theta\ \sin\theta]/(1-\omega^2)} \\ &= \frac{(1-\omega^2)^{3/2}}{4\pi}\int_0^{\pi/2}\frac{d\theta}{(1+\omega\sin(2\theta))^2}\int_0^\infty dt\,t\,e^{-\frac{1}{2}t} \\ &= \frac{(1-\omega^2)^{3/2}}{2\pi}\int_0^\pi\frac{d\phi}{(1+\omega\ \sin\phi)^2} \end{aligned}$$

To calculate the latter integral we define

$$\tilde{I}_n = \int_0^\pi\frac{d\phi}{(1+\omega\ \sin\phi)^n}$$

These integrals obey

$$\omega\frac{d}{d\omega}\tilde{I}_n - n\tilde{I}_{n+1} = -n\tilde{I}_n$$

so

$$\tilde{I}_2 = \tilde{I}_1 + \omega\frac{d}{d\omega}\tilde{I}_1 \qquad\qquad \tilde{I}_1 = \frac{2}{\sqrt{1-\omega^2}}\arccos(\omega)$$

(where we used the integral already encountered in **III**). We now find

$$I_6 = \frac{(1-\omega^2)^{3/2}}{2\pi}\tilde{I}_2 = \frac{(1-\omega^2)}{\pi}\arccos(\omega) - \frac{\omega\sqrt{1-\omega^2}}{\pi} + \frac{\omega^2}{\pi}\arccos(\omega)$$

VII: $I_7 = \langle |x|\,|y|\,\theta[-xy]\rangle$

$$I_7 = \int_0^\infty\int_0^\infty \frac{dxdy}{\pi\sqrt{1-\omega^2}}xy\,e^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

We use the relation

$$xe^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)} = -(1-\omega^2)\frac{\partial}{\partial x}e^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)} - \omega y e^{-\frac{1}{2}[x^2+y^2+2xy\omega]/(1-\omega^2)}$$

to give us, using **VI**:

$$I_7 = \frac{\sqrt{1-\omega^2}}{\pi}\int_0^\infty dy\,y\,e^{-\frac{1}{2}\frac{y^2}{(1-\omega^2)}} - \omega\langle y^2\theta[-xy]\rangle$$

$$= \frac{(1-\omega^2)^{3/2}}{\pi} - \frac{\omega(1-\omega^2)}{\pi}\arccos(\omega) + \frac{\omega^2\sqrt{1-\omega^2}}{\pi} - \frac{\omega^3}{\pi}\arccos(\omega)$$

VIII: $I_8(x) = \int dy\ \theta[y]P(x,y)$

$$I_8(x) = \int_0^\infty \frac{dy}{2\pi\sqrt{1-\omega^2}}e^{-\frac{1}{2}[x^2+y^2-2xy\omega]/(1-\omega^2)}$$

$$= \frac{e^{-\frac{1}{2}x^2}}{2\pi\sqrt{1-\omega^2}}\int_0^\infty dy\ e^{-\frac{1}{2}[y-\omega x]^2/(1-\omega^2)} = \frac{e^{-\frac{1}{2}x^2}}{2\sqrt{2\pi}}\left[1+\mathrm{erf}\left[\frac{\omega x}{\sqrt{2}\sqrt{1-\omega^2}}\right]\right]$$

IX: $I_9(x) = \int dy\ \theta[y](y-\omega x)P(x,y)$

$$I_9(x) = -\frac{\sqrt{1-\omega^2}}{2\pi}\int_0^\infty dy\ \frac{\partial}{\partial y}e^{-\frac{1}{2}[x^2+y^2-2xy\omega]/(1-\omega^2)} = \frac{\sqrt{1-\omega^2}}{2\pi}e^{-\frac{1}{2}x^2/(1-\omega^2)}$$